

# Proceedings of the International 1991 Census Planning Conference

91-030-901

244



Statistique Canada    Statistics Canada

Canada





Statistics  
Canada

Statistique  
Canada

# Proceedings of the International 1991 Census Planning Conference

Ottawa, October 8-11, 1985

STATISTICS STATISTIQUE  
CANADA CANADA

FEB 5 1997

LIBRARY  
BIBLIOTHÈQUE

Published under the authority of the Minister of  
Supply and Services Canada

© Minister of Supply and Services Canada, 1987

April 1987

ISBN 0-662-14922-X

Ottawa

La version française de cette publication est  
disponible sur demande.



## Message from Dr. Edward T. Pryor

The October 1985 International Planning Conference on the 1991 Census marked the starting point for Statistics Canada in its planning and development for the next decennial Censuses of Population and Agriculture, to take place in June 1991.

The Census of Population is Statistics Canada's largest and most visible data collection vehicle; it is the cornerstone of the entire social statistics program and provides data for thousands of users in all parts of the country. The Census of Agriculture plays a similar role in measuring the state of Canadian agriculture. This conference represented a first attempt to define the shape of the 1991 Census and how best to address the many technical, content and policy issues that must be faced.

The participation of census-taking experts from other countries – the United Kingdom, the United States, Australia and Sweden – was of enormous help to Statistics Canada. They brought to the conference not only their considerable experience, new ideas and approaches, but a sense of shared problems and issues. Such exchanges of information and experiences are and continue to be mutually beneficial to all countries.

I would like to thank all the participants and organizers of the conference for making it a success. Through their work we established a solid foundation for 1991.

A handwritten signature in black ink, reading "Edward T. Pryor". The signature is fluid and cursive, with the first letters of the first and last names being capitalized and prominent.

Edward T. Pryor  
Director General  
Census and Demographic Statistics



# TABLE OF CONTENTS

	Page
<b>Introduction</b> .....	1
<b>Welcoming Address</b> .....	3
Ivan P. Fellegi	
<b>Opening Remarks</b> .....	7
Edward T. Pryor	
<b>Session: Census Collection</b> .....	9
Automation in the Collection Environment .....	11
Brian J. Williams	
Telephone Followup – Past, Present, and Future .....	17
John A. Kazmaier Jr.	
Extension of Mail-back into Pick-up Areas .....	23
Doug Hicks	
Panel Discussion: The 1991 Collection Environment – Three Perspectives .....	29
Confidentiality/Legal Issues .....	29
Richard Barnabé	
Highlights on the Elderly and the Immigrant Population of Interest to Collection Operations .....	33
J. Boyd Underhay	
Native Persons and Remote Areas .....	35
C. Jerry Page	
Résumé of the Question Period .....	37
<b>Session: Census Geography</b> .....	39
Geographic Support for the 1990 Decennial Census .....	41
Silla G. Tomasi	
Area Master Files – A Better Way to Serve Census Needs But How Far Should We Extend Coverage? .....	47
Joel Yan	
Karole Kidd	
Jean-Pierre Parker	
A Block Program – Yes or No? .....	59
Robert Parenteau	
Computer Systems to Support Census Geography .....	67
Gordon Deecker	
Ron Cunningham	
Karole Kidd	
Résumé of Discussion: A Comparison of American and Canadian Plans for the Census in the 1990s .....	75
Sid Witniuk	
Résumé of the Question Period .....	79
<b>Session: The Content of the 1991 Census</b> .....	81
Panel Discussion: The 1991 Content Development Process .....	83
Australian Experience .....	83
Henry Kriegel	

# TABLE OF CONTENTS - Continued

	Page
United States Experience .....	84
Susan Miskura .....	
United Kingdom Experience .....	84
David Pearce .....	
Panel Discussion: The 1991 Content Proposals .....	87
Representative of Futuresearch Inc. ....	87
John Kettle .....	
Representative of Clayton Research Associates .....	88
Frank Clayton .....	
Representative of the University of Toronto .....	90
Noah Meltz .....	
<b>Session: Census of Agriculture</b> .....	95
Mail Enumeration in the United States Census of Agriculture .....	97
Cynthia Z.F. Clark .....	
presented by .....	
Charles P. Pautler .....	
A Proposal for a Land-based Census of Agriculture .....	117
G. Oliver Code .....	
Confidentiality of Census of Agriculture Data .....	121
Rick Burroughs .....	
Mary March .....	
Résumé of Discussion .....	129
Mel Jones .....	
Résumé of the Question Period .....	131
<b>Session: Census Automation</b> .....	133
Automation Plans for the 1990 U.S. Census of Population and Housing .....	135
Peter A. Bounpane .....	
Data Capture Alternatives .....	141
Dave A. Croot .....	
Decentralizing 1990 Census Data Capture .....	145
Arnold A. Jackson .....	
On the Use of Automated Coding at Statistics Sweden .....	151
Lars Lyberg .....	
The Future for Generalized Software or - .....	
Does Software Go Bad? .....	173
Mike Jeays .....	
Résumé of Discussion .....	179
Jacob Rytén .....	
Résumé of the Question Period .....	181
<b>Session: Coverage and Data Quality</b> .....	183
Issues in Coverage Measurement and Adjustment for the United States .....	185
Howard Hogan .....	
Making Data Quality Assessment More Relevant .....	191
Richard Burgess .....	
Adjustment for Non-coverage Errors .....	195
Chris Hill .....	

## TABLE OF CONTENTS - Concluded

	Page
Address Registers: Advantages and Disadvantages .....	199
David C. Whitford .....	
Applications of an Address Register in the Canadian Census .....	207
Don Royce .....	
Résumé of the Question Period .....	217
<b>Session: The Role of Research and Testing</b> .....	219
Plans for the 1991 Census in the United Kingdom .....	221
David Pearce .....	
Major Issues in the 1990 U.S. Census of Population and Housing .....	225
Peter A. Bounpane .....	
Plans for Future Australian Population Censuses .....	231
Henry Kriegel .....	
Résumé of Discussion .....	235
<b>Closing Remarks</b> .....	237
Edward T. Pryor .....	
<b>Appendix 1 - Program</b> .....	241
<b>Appendix 2 - List of Participants</b> .....	245



## Introduction

The purpose of this manual is to provide an account of the conference on the planning of the 1991 Census. This conference took place in Statistics Canada's Ottawa offices from October 8 to October 11, 1985. The manual has been organized in such a way as to respect the order of events at the conference.

The aim of the conference was to help Statistics Canada undertake the planning of the 1991 Census, while continuing the activities connected with the 1986 Census.

Throughout the conference, the representatives of the countries invited and Statistics Canada shared their views on the census and told of their experiences. Many questions, ranging from general ones on subjects such as confidentiality of data to specific ones regarding such things as the development of automation, were raised and discussed by the speakers and participants. This meeting therefore made possible fruitful exchanges among the countries involved with censuses in the 1990s.

This manual is divided into seven main sections, one for each of the seven sessions of the conference. The order of events in the conference program is respected. Each section includes the speeches that were made during the session, as well as résumés of the question periods and discussions that followed. The question periods and discussions were summarized using tape recordings made during the conference. Some speeches were also transcribed from tapes. These are printed in italics to distinguish them from original texts supplied by the speakers.

Finally, to complete the documentation on the conference, two appendices have been included. The first is the conference program, the second the list of participants.



## WELCOMING ADDRESS

Ivan P. Fellegi

Chief Statistician of Canada  
Statistics Canada

### General Remarks

It is a pleasure for me this morning to welcome you and to officially open this conference on the planning of the 1991 Census of Canada.

This conference marks the starting point of our planning efforts for 1991. By launching our planning with this conference, we hope to stimulate this very important process. The interaction and exchange of ideas which will take place during this week will provide the catalyst for much of our work during the next five years.

The interest in this topic and the importance of this conference are reflected not only by the large number of participants from within Statistics Canada but by the presence of a number of representatives from statistical organizations in other countries including the United States, Europe and Australia. Their thinking and experiences will be most valuable in furthering our future plans for the 1991 Census. Such strong and widespread interest is a very positive sign and an encouraging start.

In a certain sense, this conference is really two conferences in one. Canada is rather unique in the world in conducting its Census of Agriculture along with the Census of Population. Many of the issues discussed throughout the conference will apply to both censuses, but there will also be a special session devoted to the Census of Agriculture.

### The Importance of this Conference

The Census of Population is our largest and most visible program and is of vital importance to the national statistical system. The census provides:

- the only source of directly comparable data for small geographic areas;
- the only source of data for sparse population groups;
- the bench-mark for many of our intercensal programs such as population estimates, monthly employment and unemployment rates, and portions of the system of national accounts;
- the main source of cross-classified socio-economic information (for instance relating region, employment, education, age and family relationships).

In the same way, the Census of Agriculture provides the only set of comprehensive, small area statistics on agriculture. Our current agriculture statistics program is totally dependent for its accuracy on the census.

There are, however, many changes occurring to the environment in which we conduct the census. There is increasing pressure to reduce the costs of taking a census, and in 1986 we have seen major changes in the manner of funding the census. A large portion of our collection and processing work-force will be recruited through a student/youth employment program, and we will be attempting to recover far more of the costs of taking the census through a new pricing structure for our products and services. I suspect these are issues we share with other countries represented here.

The technology of information processing and retrieval is also changing rapidly. While this new technology allows us to produce our data more quickly and at less cost, it is a double-edged sword. The technology also has the effect of increasing the capacity and the appetite of users for data. As

they become more sophisticated, users are also demanding higher quality data and are becoming more questioning of our methods and approaches.

There are also increasing public concerns over issues such as privacy and the linkage of large data files, and over the need for the census to ask so many questions. The situation which has developed in West Germany, far from being irrelevant to us, is of considerable concern. There will be an increased need in 1991 for us to justify the need for the census to the Canadian public.

It may seem that 1991 is still far away, especially with the 1986 Census rapidly approaching. However, Tuesday, June 4, 1991 is in fact less than 68 months from today. It is not too early to begin discussion of 1991 issues. Many other countries are well on their way towards planning their 1990 or 1991 Census. If we wish to make significant changes for 1991, it is essential that we begin to plan now, even before we have the benefit of our 1986 experience.

We recently had a very visible illustration of the importance of censuses in Canada: the 1986 Census was cancelled by the government and, in response to widespread and articulate representations received from client groups, it was reinstated. It is worthwhile to draw some lessons from this rather dramatic sequence of events.

First, the major reason for reinstatement was, without any doubt, the fact that the planned 1986 Census content was regarded by our users as highly relevant and responsive to their needs. We must keep this as the most important objective.

Second, the great many positive submissions received were not offset by any negative feedback: i.e. the census had few enemies. This is the result of having a good record on a number of key issues, such as: confidentiality protection; good public communications programs in the past; sensitivity to privacy issues; and a record of continuously diminishing per household unit costs (in constant dollars) during the last four censuses.

Third, the role of influential external spokespersons was crucial. This implies the need for a strong and special effort to keep in close communication with the major users and user groups.

### **Expectations for the Conference**

Over the next 3 1/2 days, we will be hearing about and debating many new and different aspects of census-taking. We will be addressing issues of census content, the collection environment which may exist in 1991, and the use of automation, to name just three. Through these discussions, we intend to lay the groundwork for developing our strategy for 1991. By examining new approaches, while at the same time building on our 1981 and 1986 experiences, we hope to establish a strong footing for future planning.

We also hope to make some progress towards establishing our major planning assumptions for 1991 and towards identifying our priorities for research and testing. For 1986, our major planning assumption was that of a minimum change census. As a result, we were able to maintain unit costs at a very low level by minimizing research and development. While the acid test of this assumption will come with the results of the 1986 Census, we must begin now to discuss what we think is the most appropriate approach for 1991.

Finally, we hope to generate an awareness of the 1991 Census and to encourage thinking corporately about the priority issues for successfully planning a 1991 Census. The 1981 experience has again demonstrated to us the extent to which the census is a bureau-wide undertaking. Success in 1991 will undoubtedly depend on a strong commitment from all parts of Statistics Canada.

### **International Guests**

We are fortunate to have the assistance in this process of several representatives from the statistical agencies of the United States, the United Kingdom, Australia and Sweden. Three of these countries attended our census dissemination conference, held one year ago, and we are now putting to use

much of what we learned in the planning of the 1986 Census output program. In turn, several countries have expressed to us that they too learned from that conference.

With this conference on the 1991 Census, we again hope to benefit from the international exchange of expertise and experience. We share with them the challenges of anticipating new data needs, of choosing collection methodologies and processing technologies, and of generating public interest and support for a national census. We hope that this week will be of as much benefit to them as I am sure that it will be to us, and I bid them welcome.

### **Concluding Remarks**

To conclude, I would like to emphasize that this conference provides us with a rare opportunity to bring together the views of a large number of talented people on a very important topic. There is a large depository of knowledge, experience and intuition sitting in this room this morning.

We will benefit the most from this expertise through the free exchange and sharing of ideas and opinions, and I encourage all participants to take part in the discussions. In particular, I want to encourage the younger and newer members of our staff to play an active role in this conference. It is, to a very large extent, on you that the success of the 1991 Census will depend.

I would now like to turn the conference over to its chairman, Mr. Edward T. Pryor, Director General for Census and Demographic Statistics.



## OPENING REMARKS

Edward T. Pryor

Director General for Census  
and Demographic Statistics  
Statistics Canada

### Purpose of the Conference

The stimulus to arrange the conference emerged from our international contacts. About a year ago, we attended a meeting in the United States where we started to figure the planning and the timing for the censuses of the 1990s. At that time, it became obvious to us that it was time to look ahead to the 1991 Census, even if the 1986 Census was still under development.

There is a common tendency for countries that are in a five-year census cycle to become static about the decennial census while being in the peak period of the quinquennial census. This is especially true when we are faced with a quinquennial census that is as large as a decennial census.

The chronological order of censuses leads to emphasize the coming census in terms of development, collection, processing and production, and afterwards we focus on the next census.

Such an approach could affect the decennial census in different ways, and especially concerning innovation. New issues would not have time to develop and we would be bound to simply repeat the same content as the quinquennial census. I am assuming that, for the 1991 Census, we will not be able to tolerate that kind of stand-pat approach.

For the 1991 Census, we have to expect modifications and changes to areas like technology for collection, the content of the census and the release of the data. Moreover, some modifications and changes could be implemented initially for the 1986 Census to allow testing before a complete implementation.

### International Exchanges

Last fall, Mr. Alex Martin, who is the 1986 Census Manager, organized a conference on census dissemination gathering many countries. This conference was interesting because of the exchange of ideas between different countries on the specific type of dissemination, and also because of the fact that countries involved became more conscious about the number of issues which we have in common. We realized that we do share increasingly similar problems with other countries and it would be beneficial for every country to share them.

In the past, census-taking countries have had the tendency to isolate themselves, believing that their procedures for taking the census were the best. However, despite cultural differences and variations in statistical tradition, we discovered that there were advantages in discussing and sharing certain issues such as: how to maintain comparability in data versus changes in the content over time, how to make the best use of technology, how to present data, how to meet user needs, etc. From those discoveries emerged a new phenomenon, throughout the countries, that can be referred to as a "spirit of coalition".

There is another reason for having international exchanges on census information. This is because of the increasing interest of organizations like the United Nations, the International Statistical Institution and the International Union for the Scientific Study of Population who are interested in this information. These organizations are also interested in international standards for data comparability. Their concerns are shown by the recent discussions they have had during their meetings on topics such as census methods or the cost for taking a census. Their interest should encourage us to pursue our exchange on census issues.

In conclusion, I hope we have a good conference and a good exchange of ideas. In terms of my objectives, if we have one good idea which is fruitful and new or, if we possibly develop a new approach concerning collection, geography, content, agriculture, processing or output, this conference will be a great success.

I also hope for our guests, our staff and ourselves, that we will go away with notions and approaches which would not have occurred to us if we had not participated in this conference.

## **SESSION: CENSUS COLLECTION**

Chairperson:     John Riddle  
                      Regional Operations  
                      Statistics Canada

Tuesday, October 8, 1985



# AUTOMATION IN THE COLLECTION ENVIRONMENT

BRIAN J. WILLIAMS

MANITOBA AND SOUTHERN SASKATCHEWAN REGIONAL OPERATIONS  
STATISTICS CANADA

## Introduction

The challenge of census-taking in the 1990s will centre on a reduction in labour-intensive activities and increased pressure from users for more timely data. Technological advances will contribute to our ability to collect data, train personnel, control and monitor field and processing activities, and foster respondent relations. This paper will present a number of areas where various technological advances can aid collection activities.

At the same time, technological advancements will further unsettle the already complex environment in which we must fulfil our mandate. The invasion of privacy and privacy protection issues currently before us will become even more prominent as information processing technologies become more sophisticated. The 1984-85 Annual Report of the Canadian Privacy Commissioner features, on the front cover, a masked thief stealthily unlocking the chains to a micro-computer. The increased use of technology can certainly help to reduce the cost of census-taking, but will also bring a new set of respondent problems and issues.

## Data Collection

The most labour-intensive activity in the census process is the field collection operation. Collection operations will employ some 38,000 Census Representatives in the 1986 Census of Canada. Any new technology that can be applied in this phase will potentially yield significant savings.

One area that merits study is the use of a Centralized Edit and Telephone Follow-up procedure in urban centres. An internal report by Survey Operations Division in May 1983 concluded that a Centralized Edit and Telephone Follow-up would not produce a cost-saving, but in fact cost an additional \$100,000.<sup>1</sup>

This study presumed the manual Census Representative (CR) edit and follow-up would be replaced by a manual centralized edit and follow-up. If, however, the edit process could be automated through the use of Optical Character Recognition (OCR) equipment, Centralized Edit

and Telephone Follow-up would be feasible. A questionnaire designed to facilitate OCR coupled with a Computer-assisted Telephone Interviewing (CATI) application for "fail-edit" documents could yield several benefits: speedier processing, automated record keeping of returns, controlled follow-up. It must be emphasized that any major changes to collection procedures such as this must be thoroughly studied and field tested. The effect on data quality must be carefully considered. If, as well, non-response households were initially contacted through a centralized follow-up process, further cost-savings could be potentially significant.

Telephone Follow-up of non-respondents through the same CATI application is also feasible, particularly in light of recent studies of the Labour Force Survey. The Labour Force Survey (LFS) is Statistics Canada's largest ongoing survey, contacting some 55,000 dwellings monthly. The LFS has recently extended telephone interviewing into rural areas previously done in person. As a result, there has been a number of studies conducted.

Preliminary results from the first two months of the Labour Force Survey's Telephone Contact Study indicate a 70% tracing rate. In this study, newly selected dwellings are matched by address to a current telephone company subscriber file to obtain name and phone number. The household is then contacted by phone to complete the Labour Force Survey. Clearly, this technique is applicable only to large urban centres where drop-off records will show civic address. Nevertheless, in larger Regional Office centres only (i.e. Montréal, Toronto, Winnipeg, Edmonton, Vancouver) there are potentially 325,000 households that could be followed up by telephone assuming 85% mail response and 70% tracing rate. Once again, thorough field testing is essential. The consequences of any such changes are enormous and could ripple through the system.

Another area that could be explored for collection operations in 1991 is the use of hand-held computers for those areas that have traditionally been difficult to enumerate, for example, follow-up in urban core areas. This would expedite processing and with instant editing, reduce errors

<sup>1</sup> Statistics Canada (1983a).

and the likelihood of a return visit. As well, it would provide for stronger control over the follow-up process.

The potential uses of this particular type of equipment (hand-held computers) bring into focus a number of important issues. As an organization, we have been reliant on a "hard copy" record of the document or questionnaire. A decision to employ hand-held computers (or for that matter automated edit with CATI follow-up) represents a basic shift in approach. What are the implications of losing the "hard copy"? As well, there are currently some technical concerns with hand-held computers today. Those that are truly hand-held have limited storage and display. Those that have overcome these limitations are portable but not hand-held for easy use in the field, and cost considerably more. Undoubtedly these technical difficulties will be overcome in the future. But, given the lead time required to design, test, evaluate systems and procedures, decisions must be made early in the census cycle. The census, by its very nature, is not the vehicle for research and development. The emphasis must be on proven technology. The use of technology is very alluring and can often be seen as a panacea. Our expectations can outstrip current technology.

Technology can also be of benefit in some ancillary collection operations. In Western Canada the high incidence of non-resident agricultural operators results in the generation of some 50,000 Forms 6D, Agriculture Land Referral Forms (ALRF). Census Representatives are asked to identify each agricultural holding in their Enumeration Area (EA) and also to determine the operator of each holding. When the CR determines that the operator resides outside the EA, or is unable to determine the operator, a Form 6D is created. Using these Forms 6Ds, the operator is traced and a Census of Agriculture Questionnaire completed. Frequently, however, details of the operator are very sketchy and tracing grinds to a halt. The ability to access automated provincial property tax records using legal land description to identify the owner and thence the operator would be most helpful.

As public institutions computerize their records, the potential benefit to collection operations is tremendous. Access to other agencies' data banks for the Reverse Record Check, for example, would undoubtedly reduce costs and improve timeliness. This technique, whilst efficient and effective, presents a legal problem. The Privacy Commissioner in his Annual Report of 1984-85 describes computer matching or linkage as a "far-reaching, insidious threat to the way our society thinks and works".<sup>2</sup>

<sup>2</sup> Privacy Commissioner (1985).

Finally, two of the options put forth previously imply a mix of technologies for both collection and processing. Traditionally we have used a single approach to collection and processing. Perhaps in the future mixed technologies could be employed. For example, the use of OCR equipment may only be possible with the Form 2A (short form) because of questionnaire size. The Form 2A could be processed using OCR; the Form 2B (long form) in the conventional manner by keying.

One of the organizational issues that will be created through the increased use of technology is the blurring of the distinction between collection and processing. Currently, processing and collection are seen as two distinct phases. The use of hand-held computers and automated centralized edit will meld collection and processing into one operation. Collection will be redefined as the creation of a "clean" data file.

## Training

There will be 32,956 self-enumeration Census Representatives in 1986. They will receive 8 hours of self-instruction blended with 6.5 hours of class-room instruction. The cost of this instruction will be 2.47 million dollars. Travel to training classes will bring the total cost to 3.1 million dollars. Expressed in other terms, each minute of the planned 1986 CR training program will cost \$2,800.

Currently, the training of collection staff is almost exclusively a "paper" exercise. There are several areas where the use of relatively common equipment could be used to reduce costs. During the 1980 U.S. Census, a series of audio cassettes was used to introduce new Regional Office staff to the census process. A similar program could well be introduced to orient Census Area Managers (CAM) and introduce preliminary activities. Given the widely dispersed CAM staff, any reduction in Regional Office training will yield savings in travel. Audio cassettes are used extensively in our training for ongoing programs at the junior levels (e.g., Labour Force Survey, National Farm Survey). Their contribution to the effectiveness of these programs is generally recognized. Cassettes could well be used in the census program to emphasize key points. Or perhaps, brief "refresher" courses for CRs could be put on cassette.

Installed in a phone answering device, a phone number could be given during training and the "refresher" program accessed when the Census Commissioner is unavailable.

Perhaps the most dynamic home entertainment sector in the last few years has been Video Cassette Recorders (VCRs). The number in Canadian homes doubled between May 1983 and March 1984.<sup>3</sup> Today they are readily available for rent at corner convenience stores. VCRs were used on an ad hoc basis in 1981 to train Census Area Managers (CAMs) and Census District Managers (CDMs) on Census Commissioner (CC) hiring and public relations duties. The availability issue has precluded more general use. As well, a well done video tape requires considerable lead time to prepare. Video tape also is an inflexible medium. A minor procedural change could undo a lot of work. Nevertheless, some aspects of the census program could be put on video tape. Video Cassette Recorders could be put to use to train staff on the handling of refusals or difficult respondents, mapping responsibilities, and other procedures that are less effectively covered through "paper" exercises. One of the potential problems that may arise with the high percentage of students that will be hired in 1986 is a much higher than normal turnover rate of CRs. To avoid tying up key staff in a never-ending cycle of training classes, it would be advantageous to have drop-off and pick-up training on video cassettes. The video could interact with a self-administered study guide to train the replacement CR.

With the cooperation of community cable channels or educational stations, video cassettes could be used on a larger scale. Ninety-eight per cent of Canadian homes have at least one black and white television.<sup>4</sup> The current self-study program could be supplemented with pre-recorded illustrations of key points, broadcast on community or educational channels in non-peak times.

Other areas to explore are the use of micro and/or hand-held computers to assist in the training of support staff. There is currently available a relatively inexpensive (\$250) hand-held computer designed to interact with a self-administered study guide. The memory modules interface with a microcomputer to produce data on the effectiveness of various aspects of the training package. It would be worthwhile to use this application to test the effectiveness of CR and/or CC training. Areas that are more frequently misunderstood would be highlighted, and improvements could be implemented in a future census. The measurement of training effectiveness is a difficult, and hence frequently neglected area that can be addressed through the use of this technology.

Enumerator training is a crucial phase of the census operation. The success of the training program will, to a large degree, foretell the success of the census. Technology will afford us the opportunity to reduce training cost and at the same time enhance the effectiveness of the training program.

### **The Control and Monitoring of Field Activities**

Advances in technology will greatly assist in a number of areas related to the control of field activities. Maintaining lines of communication with a large widely dispersed staff is a difficult yet essential task. A single procedural change after training or request for additional information beyond the Management Information System (MIS) requires over 41,000 phone calls to implement. During the early phase of field operations, CAMs and CDMs are frequently in travel status recruiting CCs, performing field checks, etc. Itineraries are frequently adjusted to accommodate candidates or to adjust to a developing problem. Getting messages to a CAM can be difficult. There is currently being tested an audio electronic messaging system that would greatly ease this problem. "Hello Central" works very much like the Envoy 100 or other similar electronic mail systems. Audio messages are left in a mailbox which is accessed through a touch-tone phone. It can be accessed from a rotary phone using a tone simulator. The monthly charge is one-third the cost of an answering machine.

The new generation of facsimile machines will expedite processing of MIS reports as well as reduce costs. Transmission times for some newer machines are as low as 30 seconds compared to 4 or 6 minutes on older machines. In days when 7 or 8 reports are due from up to 40 CAMs, line charges will be greatly reduced and the rollout of MIS data should be quicker. More importantly, there will be more time for analysis and reaction at the CAM and Regional Office levels.

The MIS itself is a prime target for further automation. Micros will be used in the Regional Offices in 1986. In the future, micros could be used at the Census Area Manager (CAM) level to further expedite the flow of MIS data. Use of micros at this level would also help to eliminate logical inconsistencies in reported MIS data. As well, installing micros at this level could assist the CAM in monitoring a variety of activities such as CR hiring and training, the flow of pay-claims and other financial records, and the cost of clean-up.

<sup>3</sup> Statistics Canada (1983b and 1984).

<sup>4</sup> Statistics Canada (1984).

One historical bottle-neck that should be further automated is the CR pay system. In Canada, the production and issuing of pay cheques is the responsibility of a central agency, Supply and Services Canada (SSC). Statistics Canada generates pay input documents and forwards them to Supply and Services for keying and cheque issuance. The process currently is automated at the Supply and Services Canada level, but not in the ROs. There is a cumbersome manual process in ROs to generate the input documents for SSC. This has historically led to bottle-necks and numerous complaints. Merging elements of the Census Geographic Master File (CGMF) with a CR file and appropriate record of work documents to produce a pay file for input to the SSC system would both save costs and reduce the number of complaints. As we are all aware, a single ministerial enquiry can siphon off considerable time and energy at a number of levels in the organization.

### **Respondent Relations**

New telephone technology will help to make more effective use of Telephone Assistance Service (TAS) staff. With some minor regional differences, current TAS specifications call for a series of single-line sets linked on a "ring-down" or call forward system. Call distribution is not possible without the addition of an expensive piece of hardware. The same holds true for sequencing and stacking. Because of the short duration of TAS, approximately 10 days, it has not been practical to acquire this additional equipment. Automatic Call Distributor (ACD) is a standard feature of the fifth generation centrex system which will be installed in most centres in the next few years. ACD coupled with the "stacking" and sequencing of waiting calls will ensure maximum use of TAS staff. On-line monitoring systems will be in place to provide more detailed information than currently available so that refinements in the scheduling of staff can be made to better serve the public.

Another possible area to study is the feasibility of establishing a call-in system of reporting census

data. Utilities use this system to record billing information when the meter reader fails to find someone home. Clearly there are a number of problems with this idea, but some form of call-in service to at least record appointments or suggested hours to call could produce savings in personal follow-up.

### **Summary**

Increased pressures to reduce costs and produce more timely data will force Statistics Canada to examine technological alternatives to labour intensive activities. When applied to collection activities, technology may reduce costs and improve timeliness. The effectiveness of the training program can be improved through the use of technology. The ability to coordinate and direct a large field staff will be improved. On the other hand, concerns over privacy and privacy protection will remain, likely even increase, as more sophisticated equipment is employed. Technology can be very alluring. We must be certain that any new technology used has a proven track record. The "one-time" nature of a census does not allow for research and development of untried systems. As well, there is considerable "lead time" required to acquire and develop new hardware and systems. Some field testing will be required to assess the effect on data quality.

The areas that merit serious consideration for the employment of technology in 1991 are training and the use of hand-held computers. Present training practices rely exclusively on traditional "pen and paper" exercises. This is both costly and less effective than a training package incorporating new media. Each one hour reduction in the training program will yield savings of \$168,000. The elimination of one training class would yield \$846,000 in fees and expenses. The use of hand-held equipment will expedite both collection and processing as well as strengthen control over the operation.

Decisions on new equipment and processes for 1991 will have to be made soon in order that their impact can be fully considered and evaluated.

---

## REFERENCES

- Barabba, Vincent P., Richard O. Mason and Ian I. Mitroff. 1983. "Federal Statistics in a Complex Environment. The Case of the 1980 Census". *The American Statistician*, Vol. 37, No. 3, pp. 203-211.
- Bounpane, Peter A., 1983. "The Census Bureau Looks to 1990", *American Demographics*, Vol. 5, No. 10, pp. 28-32, 44-48.
- General Accounting Office (United States), 1982. **A \$4 Billion Census in 1990? Timely Decisions on Alternatives to 1980 Procedures Can Save Millions**, Washington, D.C.
- General Accounting Office (United States), 1983. **The Census Bureau Needs to Plan Now for a More Automated 1990 Census**, Washington, D.C.
- Privacy Commissioner, 1985. **Annual Report Privacy Commissioner 1984-85**, Ottawa.
- Robey, Bryant, 1983. "Achtung! Here Comes the Census", *American Demographics*, Vol. 5, No. 10, pp. 2-4.
- Robey, Bryant, 1984. "The 1990 Census: A View from 1984", *American Demographics*, Vol. 6, No. 7, pp. 24-29, 46.
- Statistics Canada (unpublished), May 1983. **Centralized Edit and Telephone Follow-up Study**, Ottawa.
- Statistics Canada, May 1983. **Household Facilities and Equipment**, Catalogue No. 64-202, Ottawa.
- Statistics Canada, March 1984. **Household Facilities and Equipment**, Catalogue No. 64-202, Ottawa.



# TELEPHONE FOLLOWUP - PAST, PRESENT, AND FUTURE

JOHN A. KAZMAIER JR.

FIELD DIVISION  
U.S. BUREAU OF THE CENSUS

## Introduction

As part of the 1980 decennial census and 1985 test census data collection plan, an edit was performed on mail returns for clarity and completeness of response. Questionnaires judged to be incomplete were followed up by telephone or personally visited so that missing information could be obtained.

Our goal in both censuses was to reduce the field costs without reducing data quality. Telephone followup is a means to reduce the amount of personal visits, and consequently, reduce costs. We will continue to conduct telephone followup during the upcoming censuses.

## Experiences from the 1980 Census

### Census Methodology for Edit/Followup System

- Mail return questionnaires and enumerator completed questionnaires were clerically edited.
- Questionnaires identified during edit as incomplete were sent to either telephone followup or personal visit followup.<sup>1</sup> This included long form questionnaires with 20 or more incomplete questions and short form questionnaires with four or more incomplete questions. Telephone followup proved to be more efficient for this group of incomplete questionnaires.
- The telephone followup clerks were trained to contact the household and re-ask the questions identified by edit as incomplete. If the respondent had not entered a phone number on his or her questionnaire, the telephone followup clerk was instructed to use telephone directories and "criss cross" directories to obtain the phone number. As a last resort, the clerk could call directory assistance for the phone number. If the clerk was unable to contact the household after five attempts or contact was made with an unacceptable respondent, the questionnaires were set aside for personal visit (PV) followup.

- Quality control clerks verified that either every question marked during edit had been satisfactorily completed by the telephone clerks or that the questionnaire had been marked for PV followup. If a PV marked questionnaire had less than four unacceptable items (less than 20 for long forms), the case was marked as complete.

## Staffing and Timing

Approximately 4,000 telephone lines were budgeted for telephone followup nationally. Approximately 950 of these lines were previously used for the telephone questionnaire assistance operation. Telephone followup was conducted from 87 centralized offices nationwide in 1980.

In 1980, centralized telephone followup began approximately one week after the onset of the edit operation. It was completed about one week after the edit was completed.

The telephone operation was functional six days a week, Monday through Saturday.

Employees worked the following times in centralized offices:

### Day Shift

#### 1. Office Operations Supervisor

Monday through Friday - 8:30 AM - 5:00 PM  
Saturday - 8:30 AM - 1:15 PM

Either the Office Operations Supervisor or Office Operations Assistant worked Saturday afternoons depending on the preferences of the individuals involved.

#### 2. Senior Office Clerks

Monday through Saturday - 8:15 AM - 3:15 PM (with a half hour for lunch)

<sup>1</sup> In centralized district offices, incomplete questionnaires were followed up by telephone from the central office. In decentralized district offices, incomplete questionnaires were distributed to field enumerators, who contacted the households by using their home phones or by conducting personal visits.

3. Quality Control Clerks  
Monday through Saturday - 8:15 AM - 3:15 PM (with a half hour for lunch)
4. Telephone Followup Clerks  
Monday through Friday - 8:30 AM - 3:00 PM (with a half hour for lunch)

### Night Shift

1. Office Operations Assistant  
Monday through Friday - 1:15 PM - 9:45 PM  
Saturday - 1:15 PM or 5:00 PM - 9:45 PM (depending on hours worked by the Office Operations Supervisor)
2. Senior Office Clerks  
Monday through Saturday - 2:45 PM - 9:45 PM (with a half hour for lunch)
3. Quality Control Clerks  
Monday through Saturday - 2:45 PM - 9:45 PM (with a half hour for lunch)
4. Telephone Followup Clerks  
Monday through Saturday - 3:00 PM - 9:30 PM (with a half hour for lunch)

### Evaluation Results

An evaluation was completed to determine the effectiveness of the 1980 edit/followup system. These findings were made for the telephone followup on mail return questionnaires:

1. In centralized offices, telephone followup resolved only 26 per cent of the population and housing questions marked for followup by edit clerks.<sup>2</sup>

<sup>2</sup> The base used in determining this rate includes incomplete questionnaires that were judged acceptable based on edit tolerance rules.

2. Telephone followup was much more successful in resolving 100 per cent questions than sample questions. The success rate for 100 per cent population questions was 37.2 per cent, while the rate for sample population questions was 27.2 per cent.
3. The impact of telephone followup in centralized offices was less than expected, although still better than the impact of personal visit followup. The impact is illustrated below:

### ITEM NONRESPONSE

<u>Incoming</u>	<u>After Clerical Edit</u>
23.0%	16.9%
<u>After Telephone Followup</u>	<u>After Personal Visit Followup</u>
12.1%	10.0%

The conclusion drawn from this analysis is that PV followup did not reduce the item nonresponse substantially after telephone followup.

Even though telephone followup reduced the item nonresponse by about 25 per cent, those questions with the highest incoming nonresponse rate were not necessarily reduced by the full 25 per cent for item nonresponse.

4. It costs approximately \$1 to follow up a questionnaire by telephone as compared to \$4 if it had to be returned to the field.

### Experiences from the 1985 Test Censuses

#### Census Methodology for Edit/Followup System

- Mail return questionnaires and enumerator filled questionnaires from the collection sites (Jersey City, New Jersey and Tampa, Florida) were processed at our permanent processing site in Jeffersonville, Indiana.
- A computer edit identified those questionnaires which failed edit.
- A clerical review unit attempted to repair the failed edit questionnaires.

- Mail return questionnaires that could not be repaired were sent to a telephone unit for followup.
- The telephone followup clerks were trained to contact the respondents and re-ask the questions identified by edit as incomplete.
- If a telephone clerk was unable to make contact after three attempts or if a respondent requested a personal visit, the original mail return questionnaire was returned to the data collection office for personal visit followup.

### Staffing and Timing

The telephone followup operation lasted about two weeks.

Only one work shift was in operation. The shift ran from 11:30 AM until 8:00 PM daily (Monday through Friday).

The following employees were assigned to the work shift:

1. Primary Supervisor
2. 2-3 Lead Operators (Assistant Supervisors)
3. 55 clerks (even though 60 telephones were installed, they were not fully utilized)

### Preliminary Evaluation Results

No formal evaluation has been conducted for the telephone followup for the 1985 test census. However, there are some preliminary results:

1. Approximately 80 per cent of the failed edit cases for mail returns had telephone numbers (either written by the respondents on the questionnaire, or obtained by a clerk in Jeffersonville from telephone directories or directory assistance).
2. Telephone followup clerks contacted 78.7 percent of the households whose questionnaires failed edit.

(NOTE: Most of the telephone followup clerks were new hires. A few experienced telephone interviewers were also used.)

### Potential Benefits of Telephone Followup

1. Past experience shows that the unit cost is considerably lower for telephone followup than it is for personal visit followup on failed

edit cases. This is one of the primary considerations for using the telephone for as many of the failed edit cases as possible. In the 1980 decentralized offices, failed edit cases were sent to enumerators. Enumerators were instructed to use their home telephone to resolve those units for which a phone number was provided, either on the questionnaire or in a phone directory.

2. With the emphasis for using the telephone as much as possible, the Census Bureau will also test methods for conducting nonresponse followup by telephone in the 1986 test censuses, as well as failed edit resolution.

Nonresponse followup is the operation done to contact and interview those households who did not return a completed questionnaire in the mail. The census enumerators will be provided with assignment lists containing all the addresses for households requiring a questionnaire. Some of these nonresponse addresses will also include telephone numbers. The enumerators will be trained to use the phone numbers to obtain complete interviews. If a unit is done by telephone, the enumerator will indicate it was a telephoned case on the questionnaire.

Some telephone numbers will be suppressed from the nonresponse followup assignments. These cases will be used as a control group for data comparisons to those units completed by telephone. Also, measurements of coverage (where coverage means enumerating the correct persons in the correct place without duplication) will be made by personal visits reinterview results.

Also, plans are being made to evaluate the "terminal telephone", where some of the failed edit cases will never be returned to the field for personal visits. Instead, they will be sent back for further telephoning. Other failed edit units will be reassigned for personal visits. The item nonresponse will then be compared, along with costs of additional telephoning versus personal visit costs.

3. The Census Bureau is using computer assisted telephone interviewing (CATI) for some of its one-time and current surveys. A CATI facility has been established in Hagerstown, Maryland.

If the Census Bureau pursues a computer assisted telephone followup of failed edits on the census, it may be possible to further

reduce the number of nonresponse items. A computer assisted program forces the terminal operator/telephone clerk to input new data for unresolved items, without thumbing through a paper document to find all the items marked for followup. It also allows for the design of specialized probes. Thus, the operational design may be more desirable under a computer assisted program.

Also, an automated call management could be utilized to allow for more efficiency in call-backs. The system would have a database of all failed edit cases, but it would only allow cases to be worked at certain hours of the day, based on the record of previous contacts.

4. The 1980 evaluation indicated that the impact of telephone followup in reducing the percentage of nonresponse items was greater than the impact made by enumerators who personally visited households to resolve failed edits. Thus, our conclusion is that telephone follow-up is a worthwhile operation.

#### Logistical and Operational Difficulties

Some of the problems listed below have been identified in earlier censuses; others are anticipated.

1. Many respondents are unwilling to enter their phone numbers on the census questionnaire. Also, there has been an increase in the number of unlisted or unpublished phone numbers in certain areas.

Public suspicion seems to be on the increase, and more and more persons do not wish to give information about themselves or their living quarters over the telephone. More private companies are using the telephone for contacts for product sales and promotions. Many respondents are "turned off" by telephone sales persons. In some cases, the Census Bureau telephone interviewer is perceived as a telephone "solicitor".

2. With the AT&T divestiture, there could be considerable difficulties in installation of local or WATS Lines in a central location, particularly if the telephone followup is highly centralized.

With deregulation, phone lines and instrument installation costs may be much higher in centralized locations than they were in 1980.

3. If a telephone "look up" operation is required, and telephone and criss cross directories are the only reference sources, considerable time could be spent locating the correct phone number. It is extremely difficult to locate phone numbers for persons living in multi-unit structures, particularly when they have no identifiable unit designation (such as Apt. 1 or Apt. A). Considerable clerical time is required to match names and addresses. Computer searching may be a viable option, but extensive research is required for computer address matching.

4. The 1980 census experience indicated that the work flows were uneven during the telephone followup operation. Telephone clerks ran out of work, but the previous operations (edit and edit QC) sometimes did not generate additional work at the right times for the telephone staff. This affected the costs, because of inefficiencies in assigning work, and keeping everyone busy at all times. However, with an automated approach (computer edit/CATI), this problem is nonexistent.

#### The Census Bureau's Objectives/Goals for Successful Telephone Followup

1. The Census Bureau is committed to conducting the census quickly at a relatively low unit cost while maintaining a high level of data quality. With the budgetary and time constraints, we will need to use telephone followup where practical. Any telephone followup operation must be designed to allow for maximum success in reducing the nonresponse items on the questionnaires failing edit.

One possible way to increase the chances of success is to design the questionnaire for ease of use. Some telephone clerks could not readily identify the items marked for followup, because of the design of the 1980 census questionnaire itself.

2. It makes good sense to resolve failed edit units from the same location where mail return questionnaires are received. If the actual document is used to contact the household and resolve failed edit items, then it is logistically and operationally practical to have a centralized telephone followup conducted on mail return questionnaires. Otherwise, it would be necessary to distribute failed edit questionnaires to several hundred data collection

offices (and to several hundred employees) for failed edit resolution. The documents would then have to be transported back to the central site for data conversion recycling.

However, it may not be possible to install several hundred telephone lines in one location for telephone followup, because of the lack of telephone lines to a specific building. Also, some telephone exchange areas are limited to the number of lines.

3. Computer assisted telephone principles, such as an automated call management system, could be implemented to increase operational efficiency.

However, it may not be cost effective to install a network of computer terminals for each and every telephone clerk. In 1980, there were 7,500 telephone clerks required for telephone

followup. Since the operation ran for only 3-4 weeks, it probably would not be practical to have several hundred computer terminals linked to a mini computer(s) for conducting computer assisted telephone interviewing.

An automated call management system would have to be tested for feasibility. If CATI is not in place, it may be very difficult to have automated call management. Data would have to be keyed from call records, and then someone would have to manually assign unresolved failed edits to various work shifts, based on the results of the automated call manager sorting. The automated call manager would probably be personal computer-based.

If we have an automated field data collection system with hand-held computers, then the call manager approach may be viable.

#### **Note:**

#### **1986 U.S. Test Census**

*During his presentation, Mr. Kazmaier elaborated further on the 1986 Test Census that will be conducted in Central Los Angeles, California, with approximately 240,000 housing units and in eight counties in East Central Mississippi, with approximately 80,000 housing units.*

*In both places, the edit and follow-up methodology will be similar. Mail return questionnaires that cannot be completed in a telephone follow-up will be returned to the collection office for personal follow-up.*

*The 1986 Test Censuses are not only used to test telephone follow-up for cases of failed edits but also to test telephone follow-up for non-response households.*

*A list of telephone numbers for the test area in Los Angeles has been matched to the address control*

*file. This matching is done using the Geographic Base File (GBF) which is a machine-readable rendition of street maps. The enumerators will be instructed to use the telephone list to follow up non-response households.*

*In Mississippi, a match between telephone numbers and the address control file was impossible because the GBF does not cover the test area. In this case, the enumerator will have to first make a personal follow-up visit and if unsuccessful will then attempt a telephone follow-up, having obtained the telephone number through such means as neighbours and telephone books.*

*Finally, the evaluation of the 1986 Test Censuses will help to determine if the use of telephone follow-up for failed-edit questionnaires and for non-response households affects data quality and coverage.*



# EXTENSION OF MAIL-BACK INTO PICK-UP AREAS

DOUG HICKS

SURVEY OPERATIONS DIVISION  
STATISTICS CANADA

## Introduction

The current census employs self-enumeration methodology to enumerate 99% of all households and agricultural holdings in Canada. There are two forms of self-enumeration: mail-back,<sup>1</sup> which is used in larger urban areas, and pick-up,<sup>2</sup> which is used in small urban areas and rural areas. This paper describes the implications of adopting the universal use of mail-back in 1991. Potential cost efficiencies, reduction in labour intensive activities, operational concerns, assignment creation, the impact on the Census of Agriculture, and other considerations are addressed.

## 1. Cost Efficiencies/Labour Intensive Activities for the Census of Population

The table below demonstrates savings that could be realized if a mail-back methodology was employed in pick-up areas with no change in assignment size in 1986.

There is every reason to believe that as we approach 1991, savings by converting from pick-up to mail-back would be even greater. Labour and travel cost increases associated with a pick-up methodology are expected to exceed any postal increases associated with a

**Table 1. Comparison of Hypothetical Savings Between Mail-back and Pick-up Methodology, Canada, Census of 1986**

Geographic designation		M/B rate	P/U rate	*Savings per hhd	Estimated hhlds	Savings
		(\$)	(\$)	(\$)		(\$)
A	2A 2B	1.30 1.98	1.71 2.36	.032	960,000	31,000
B	2A 2B	1.43 2.12	1.88 2.53	.070	230,000	16,000
C	2A 2B	1.66 2.38	2.19 2.84	.144	400,000	58,000
D	2A 2B	2.00 2.75	2.64 3.28	.246	850,000	210,000
E	2A 2B	3.17 4.04	4.18 4.82	.592	285,000	170,000
						485,000
*Average savings per household by geographic designation are determined as follows:						
$((2A \text{ P/U rate} - 2A \text{ M/B rate}) \times 4 + (2B \text{ P/U rate} - 2B \text{ M/B rate})) - ((2A \text{ postage rates} \times 4 + 2B \text{ postage rate}) \times 90\% \text{ mail return rate})$						
5						

<sup>1</sup> Mail-back is defined as the return of a questionnaire through the mail to the Census Representative for each private dwelling at which a questionnaire was dropped off. The Census Representative edits returned questionnaires and conducts telephone follow-up and/or field follow-up for all edit failures and non-response households.

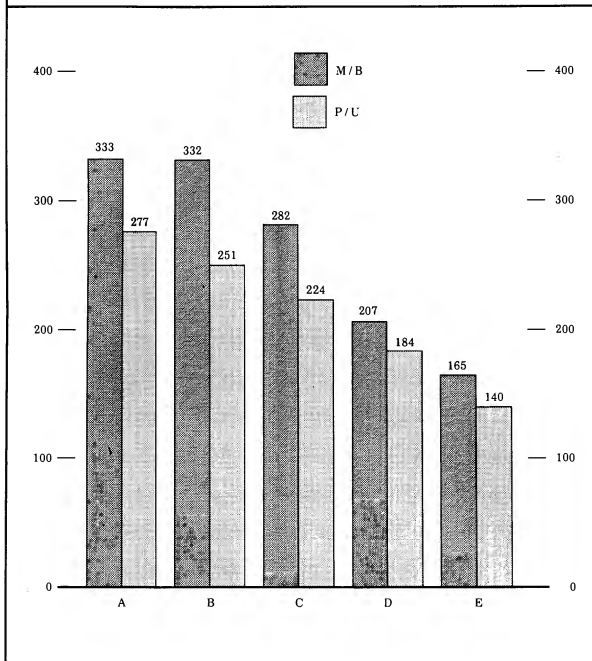
<sup>2</sup> Pick-up is defined as a return visit by the Census Representative to each private dwelling to pick up and edit all questionnaires dropped off.

mail-back methodology. A pick-up methodology also relies heavily on making contact during the retrieval phase. As the non-contact rate increases, the number of return visits required also increases resulting in higher retrieval costs. This problem, and its

associated costs, would be minimized with a mail-back methodology.

The chart below compares average assignment size in 1981 by methodology for similar geographic designations.

**Figure 1. Comparison of Average Assignment Size Between Mail-back and Pick-up for Similar Geographic Designations, Canada, 1981 Census**



Dollar savings would be realized through a decrease in staff requirements if assignment sizes in pick-up areas approached those in mail-back areas. The reduction of Census Representatives has a multiplying effect – a reduction of 15-20 Census Representatives results in the reduction of one Census Commissioner, one Census Commissioner Administrative Assistant, and part of one Quality Control Technician. In 1986 a reduction of one Census Representative would result in a saving of \$150 to \$200; a reduction of one Census Commissioner would result in a saving of \$6,500 to \$7,000; the reduction of one Census Commissioner Administrative Assistant would result in a saving of \$550 to \$600; and the reduction of one Quality Control Technician would result in the saving of \$700 to \$800. If assignment sizes in current pick-up areas approached those in mail-back areas, it is expected that staff could be reduced by 1,500 persons resulting in a saving of \$750,000.

On the negative side, the universal use of mail-back would result in some additional expenses:

- delivery of questionnaires from the Census Commissioners' office to Census Representatives;
- potential long-distance charges for telephone follow-up in some areas;
- additional production costs – associated with the mail-back envelope.

In total, these expenses would not be expected to exceed \$300,000.

## 2. Operational Considerations

### 2.1 Telephone Systems

The successful extension of mail-back is dependent upon a high mail response rate and a capability of performing telephone follow-up on questionnaires which fail edit and, possibly, for non-response households. One potential restriction on telephone follow-up is the prevalence of the party-line system in rural areas.

Based on 1980/81 data, it has been estimated that the percentage of dwellings with private lines is 50% with an equal percentage of dwellings being served by party lines.

It is probable that the incidence of party lines has decreased since 1980 and will

continue to decrease as we approach 1991. However, by 1991, it is expected that the incidence of party lines in rural areas will still be significant and the potential will exist for breaches of confidentiality should current mail-back/follow-up procedures be adopted. This is one factor that must be addressed if the extension of mail-back is to be considered a viable option. It is not believed to be an insurmountable problem as there are now many surveys which make use of the telephone technique and it is probable that this trend will become more prevalent in the future.

### 2.2 Rural Postal System

The successful extension of mail-back is dependent upon the timely receipt and distribution of mail returns. Three factors can affect the rate at which Census Representatives receive questionnaires:

- public cooperation in completing and mailing the questionnaires;
- the ability of Canada Post to process returns and turn them over to Census Commissioners;
- the efficiency of Census Commissioners in sorting returns and distributing them to the Census Representative.

#### 2.2.1 Public Cooperation

Given the success of the 1981 Census and mail return rates during the 1977 Extension of Mail-back Test, there is reason to believe that acceptable levels of mail response would be achieved to support a move to extension of mail-back.

#### 2.2.2 Canadian Postal Systems in Rural Areas

Extension of mail-back areas introduces some complications in the mail return process. In most pick-up areas, the respondent cannot drop the questionnaire in the corner mailbox. It is picked up by a local mail contractor or the respondent is required to go to a local Post Office to mail it. From the local Post Office, the mail must find its way to the Post Office which serves the Census Commissioner. A small study conducted by the Post Office in Census Commissioner Districts where potential mail return problems exist indicated a

2-4 day turn-around time. The potential for a slight delay in receipt of mail returns is not one which should be considered as a reason for rejecting extension of mail-back. While inconvenient to the field operation, slight delays can be dealt with.

### **2.2.3 Distribution of Mail Returns by Census Commissioners**

It is apparent that additional expenses would be incurred in ensuring the timely distribution of questionnaires by Census Commissioners to Census Representatives. It would be necessary to evaluate current operational schedules and plans and make recommendations to minimize the cost impact. The potential of having returns go directly from the Post Office to the Census Representative could also be investigated.

### **2.3 Timeliness**

If extension of mail-back is pursued, timeliness of collection operations under mail-back must be reviewed. Although results of the 1977 Extension of Mail-back Test indicated mail-back was an acceptable alternative, it did not indicate that it was preferable to pick-up. The factors which were identified as potential problems and which contributed to a longer collection process should be studied to attempt to minimize their negative impact.

### **3. Assignment Creation**

This topic, although not restricted to extension of mail-back, is being included because, if dollar savings are to be realized, there must be extensive redelineation of pick-up areas.

Currently, a Census Representative is responsible for enumerating a geographically delineated area called an enumeration area. Because of geographic restrictions, the cost of redelineation, and the desire over past censuses to stabilize boundaries for data comparability, it is difficult to take advantage of economies that can be realized through increased work-loads.

The closer we move to desired assignment sizes, the greater are the economies to be realized in terms of staff and dollars. This is not only true for extension of mail-back, but is also applicable to any collection methodology.

A system which allows for flexible assignment creation is preferred. As efficiencies are identified in collection methodologies, the means to realize them through increased assignment size must be available.

### **4. Census of Agriculture**

The major concern with respect to extension of mail-back is agriculture enumeration. Cost, timeliness and potential coverage issues must be addressed.

#### **4.1 Cost**

Under a pick-up methodology the Census of Agriculture only incurs enumeration costs. It piggy backs on the Census of Population. Therefore, costs such as arrival time, travel time, mileage, etc., are paid for by the Census of Population. This would not be the case with extension of mail-back. It is known that, historically, the completion of the Agriculture Questionnaire is more enumerator-dependent than the Population Questionnaire. Edit failures and cases of non-response will not necessarily coincide with those for Population. Lack of information on factors such as potential mail response, questionnaire edit failure rates, success of telephone follow-up, numbers of field follow-ups, etc., make it very difficult to estimate the dollar increase which might result from a mail-back methodology. A detailed study would be required to assess the magnitude. Although there is no firm basis for the amount, previous assessments have crudely estimated the increase at \$500,000.

Consideration must also be given to the fact that conducting the Census of Agriculture with Population could minimize the savings achieved by the Census of Population by placing a restriction on assignment size increases.

#### **4.2 Timeliness**

Evidence from the 1977 Extension of Mail-back Test indicated that timeliness could

become a problem in areas of high agricultural concentration particularly where there is a significant number of non-resident operators.

#### 4.3 Coverage

The 1977 Extension of Mail-back Test also indicated that there was potential for a higher undercoverage rate in agricultural areas not covered by township plans.

The above concerns must be studied if the mail-back methodology is to be considered an acceptable alternative for the Census of Agriculture.

The following is a short list of items that could be explored:

- simplify and redesign the Agriculture Questionnaire to make it easier to complete, which could result in increased mail-response and reduced edit failures;
- relax Agriculture Questionnaire edit steps;
- develop a publicity program emphasizing the mailing back of the Agriculture Questionnaire;
- study the success and failures of other agricultural surveys, particularly the National Farm Survey which is now experimenting with a mail-back/telephone follow-up approach;
- provide enumerators with lists of non-resident operators;
- include a filter question on the Population Questionnaire "Do you operate an agricultural holding?"

It is also suggested that the possibility of not conducting the Census of Agriculture at the same time as the Census of Population be investigated. The Census of Population could identify agricultural operators, and a mail-out/mail-back approach could be used with field follow-up of both non-response cases and those not resolved by telephone. This approach would allow the Census of Population to maximize savings through assignment creation and would, as well, virtually eliminate any potential local enumerator problems with respect to the Census of Agriculture.

## 5. Other Considerations

### 5.1 Public Reaction

A motivational study conducted after the 1977 Extension of Mail-back Test indicated that there was no reason to expect respondents to react unfavourably to the use of mail-back in rural areas. Indeed, the evidence suggests the reaction would be favourable. A mail-back methodology also gives the respondent a perception of greater confidentiality and minimizes the enumerator/respondent contact.

### 5.2 Publicity

It is recognized that confusion or "spillover" effects from publicity exist when both mail-back and pick-up procedures are used. This is particularly true for residents of pick-up areas who are constantly subjected to media emanating from urban centres. This confusion would be eliminated with the use of a standardized methodology and hence a standard publicity program.

### 5.3 Standardization of Procedures

One self-enumeration methodology would reduce or eliminate the requirement of some procedural and training material and would simplify the task of census supervisors who must supervise mixed methodology areas.

### 5.4 Coverage - Population

It is not felt that coverage would be adversely affected by extending mail-back. Coverage is established primarily at drop-off, and there is no difference in the relevant drop-off procedures under pick-up and mail-back.

## Conclusion

In summary it is recommended that extension of mail-back be given serious consideration for 1991. It is recognized that the move to extended mail-back is not without risks, particularly as it relates to the Census of Agriculture, but it is felt these risks can be minimized. Regional Operations believes that the universal use of mail-back in 1991 can be made to work both efficiently and effectively, and support this strategy.



# PANEL DISCUSSION THE 1991 COLLECTION ENVIRONMENT - THREE PERSPECTIVES

## CONFIDENTIALITY/LEGAL ISSUES

RICHARD BARNABÉ

QUEBEC REGIONAL OPERATIONS  
STATISTICS CANADA

### Introduction

Legal and confidentiality issues are critical factors for a census. My purpose today is to explore with you what the future holds, or may hold, in this regard and what actions we may have to take to adapt. Some of my comments may sound provocative or even heretical because I will outline scenarios that entail moving away from what are often considered to be sacrosanct features of census-taking.

However, these aspects, which are technical in nature, may indeed have to give way to socio-political imperatives if censuses are to remain viable and acceptable undertakings in the future.

I will first discuss legal aspects, then how this legal framework fits into the broader context of public opinion and conclude with their policy or operational implications for us.

### Legal Context

The legal landscape is outlined by legislation, but its colours and contours are detailed by jurisprudence. The Canadian picture is not unlike that of most other Organization for Economic Co-operation and Development (OECD) countries: an enabling law, the Statistics Act, specifies the mandate of the census-taking agency and its obligations as well as those of its citizens. However, other legal realities directly impact upon the interpretation of the Statistics Act by the courts: the Charter of Rights and Freedom and the Privacy Act are prime examples.

These constitutional and legislative realities have led to a series of limits on the nature and extent of governmental information gathering activities. It can be said that a doctrine of "reasonable limits" and the primacy of people's right to privacy is gradually emerging, but to varying degrees depending on the country.

In Germany and the Netherlands, censuses have been postponed or cancelled because of legal challenges.

In Canada, the U.S.A., and the United Kingdom, the courts have generally upheld censuses and their compulsory nature, although isolated challenges have been upheld on the basis of legal technicalities.

In Denmark, the use of permanent registries is authorized and strongly supported by the executive and judiciary.

The general trend, however, is that a greater balance is sought between legitimate governmental requirements and individuals' rights. That is important because this area is relatively new legal ground in many countries. Therefore, the internationalization of related jurisprudence is more likely than for other domains where strong local traditions exist.

The implication for us is that we can probably expect continued legal support if we ensure that our methods and practices respect the above-mentioned balance. In practical terms, it raises with particular acuity the issues of identification, retention of identifiable records, and the integrity of our operations from a security viewpoint.

### Public Opinion Context

Legal issues are very important because they establish the framework within which we operate. However, the success of a census will always rest on the public's acceptance of its necessity and willingness to provide the requested information.

Public opinion support for the census can be illustrated as a platform which rests on three pillars:

- legal obligations;
- recognition of the necessity of the census;
- the credibility of the guarantees we offer concerning the confidentiality and use of the information collected.

These pillars help withstand the negative pressures that could cause the platform to collapse. These pressures can be crystallized as the "Orwellian Syndrome": a concern over the government's habit of obtaining information on everyone to better control all aspects of their lives. This syndrome can be exacerbated or alleviated depending on the public's view of the government's effectiveness in fostering the social and economic well-being of its citizens. However, the census usually raises three concerns in this regard, regardless of the prevailing mood:

1. Concern over the amount of information available to government.
2. Concern over the use made of the information.
3. Concern that information gathering is a breach of peoples' right to privacy.

A successful census cannot be taken unless these concerns are kept to manageable levels.

### The Future Context

Before addressing how we can deal with these concerns, we must assume the likely environment in which censuses will be conducted. I submit that the following conditions are likely to prevail:

- Our society will be increasingly litigious, and legal challenges of the census will be more frequent.
- Privacy, and the right thereto will remain a forefront issue.
- Information technology will have made considerable progress and its potency (both as a benefit and a threat) will be even more generally recognized.
- Societal issues and increasingly sophisticated special interest groups will augment the need for detailed, timely and reliable data.

These conditions represent opportunities and potential problems for us. It is important that we define appropriate responses now.

### Conclusion

The responses must deal with the following issues:

#### 1. Confidentiality

We must offer credible guarantees relative to three types of potential threats:

#### Direct threat:

Willful or accidental release of confidential information by an employee. It is the most easy for us to cope with because we already have an appropriate legal framework. We must, however, guard against the temptation of achieving efficiency or effectiveness gains at the expense of utmost prudence in this matter. This is an area where cost reduction or operational expediency must not be recklessly pursued.

#### Indirect threat:

Dangers associated with the existence of separate files which are individually secure but that could yield confidential information by merging or matching them. Techniques such as random rounding, suppression and restricted access already exist to circumvent this possibility. Our vigilance must remain extreme.

#### Perceived threat:

The question of what could be done, now or in the future, with the information by an ill-intentioned (in the eyes of the public) government. This fear is difficult for us to alleviate as long as we collect and retain identifiable information. Our institutional record has been our best advocate up to now, but it may not prove adequate enough in the future given the emerging social attitudes.

## 2. Identification

If the collection and retention of identifiable information remains a major concern despite the confidentiality guarantee, more drastic measures must be envisaged. Not collecting names or destroying all traces thereof early in the operational process have significant implications on concepts such as family formation, on questionnaire design, on follow-up techniques and on coverage measurement. However unappealing these implications may be, they must be assessed and alternate approaches explored, if only as a precaution against currently unforeseen external pressures that could force us in this direction. The last thing we need is to be coerced unprepared up that avenue.

### 3. Legislation

Should statistical agencies openly recommend the tightening of laws on information to reassure the public of the harmlessness of providing information? For example, the laws could more severely restrict information sharing, the merging of files, or the collection, or at least retention, of identifiable records. Some of these restrictions may hamper our technical possibilities, but they may become essential to the continuation of our activities.

On the other hand, we may want to advocate methods such as permanent registries constructed from existing (and presumably not too contentious) records, and move away from direct data collection. Again, the legal and political implications are far-reaching.

The basic question that needs to be addressed now is whether we want to take a pro-active role in changing the status quo.

### 4. Communications

It is the sincere belief of census-taking organizations that censuses are cost-effective national endeavours. However, our efforts to underscore that fact and to thus generate widespread public support are irregular at best, and tend to be centred around census days. It may be well advised to re-examine our approach in this regard and to invest more resources in this area, albeit, in these times of restraint, at the expense of technical pursuits which are dearer to our professional hearts but perhaps less critical to our long-term strategic interests and that of the public we must serve.



# HIGHLIGHTS ON THE ELDERLY AND THE IMMIGRANT POPULATION OF INTEREST TO COLLECTION OPERATIONS

PRESENTED BY: J. BOYD UNDERHAY

NEWFOUNDLAND AND LABRADOR  
REGIONAL OPERATIONS  
STATISTICS CANADA

## Elderly Population

The elderly population tends to be more difficult to enumerate for a whole host of reasons such as:

- general mistrust of officials;
- confusion over government forms;
- fear of disclosing personal information.

In 1991, the elderly will be a powerful and important force. The elderly will:

- represent 11.8% of the population;
- grow at a faster rate than our youth;
- live longer;
- be better educated;
- be less dependent on government transfers;
- tend to live in private households as opposed to institutions;
- tend to live in small urban centres.

Two particular phenomena are of particular interest to enumeration of the elderly in 1991:

### 1. Early retirement:

- There is a possibility that the early retirement trend of the 1980s will continue into the 1990s.
- If the 55-64 age group (i.e. early retirees) is added to the elderly (65+), this overall group would account for an estimated 20.8% of the total population.
- As a possible "special interest population", the importance of this group would equal that of transient youth in the 1980s.

### 2. High proportion of immigrants among the elderly:

- Immigrants constituted almost 1/3 of the elderly population in 1981.

- If this portion increases in 1991, two characteristics of this population have to be considered in collection planning:

- (a) tendency to retain the first language,
- (b) concentration of this population in high-rise buildings in large metropolitan areas, especially in Ontario, British Columbia, Quebec and Alberta.

## The Immigrant Population

The immigrant population overall tend to be more difficult to enumerate. The factors complicating enumeration are the same as for the elderly population but, in addition, a language problem may exist.

The characteristics of the immigrant population of particular interest in 1991 enumeration are:

- a possible increase in the size of the immigrant population;
- a tendency to be better educated than the general population;
- language problems (in 1981, a small minority of immigrants were unable to speak either of the two official languages);
- a tendency to have a large proportion of investment income;
- a tendency to be more dependent on private pensions;
- a tendency to live in high-rise buildings;
- a tendency to live in multiple-family dwellings.

Also, there has been a noticeable increase in the number of Asian immigrants over the past years. This group will eventually form part of the elderly population. This will in effect be reflected in the social characteristics of the elderly population of the 1990s.



# NATIVE PERSONS AND REMOTE AREAS

C. JERRY PAGE

## ALBERTA, NORTHERN SASKATCHEWAN AND N.W.T. REGIONAL OPERATIONS STATISTICS CANADA

### Introduction

Both remote areas and the native population are of concern to census takers. The native population live mainly in remote areas, and these remote areas are mainly inhabited by the native population.

### Remote Areas

In Canada, remote areas are comprised of the Northwest Territories (N.W.T.), the Yukon, the coast of Labrador, and the northern part of most provinces.

This presentation will focus on the N.W.T., which represents one-third of the surface area of Canada. It has a scarce population of only 46,000 persons, comprised of 16,000 Inuit, 9,000 status Indians, 5,000 Métis and non-status Indians, and 16,000 persons of European descent. There are 63 communities, six of which make up 48.1% of the total population. Of these, 57 communities have an average population of approximately 400 persons. The main language is English, but Inuktitut (Inuit), French and Slavic are also spoken.

Communication and air transportation are satisfactory, with the exception of very small settlements.

Two enumeration systems were used in the 1981 Census. There was the regular enumeration on June 3 and an early enumeration commencing in March. The early enumeration was initiated because of costs and data quality. If we waited until spring break-up (shortly after Easter), 15% to 20% of the native population would not be counted as they left the communities to pursue their traditional way of life "out on the land". If this did occur, the undercount of the population would have increased significantly as well as the increase in costs in locating the scattered population.

There are still many problems facing census takers, such as recruiting local people who are qualified and willing to work for such a short period of time, employee turnover, and supervision and monitoring of completed tasks in outlying communities.

A proposal was made for the 1986 Census, in which a team of census takers from the South would carry out early enumeration in the North. Based on this team approach, a group of people would fly into an isolated community and conduct the enumeration within a set time frame using local enumerators and/or interpreters as required.

Our objectives for the 1991 Census are to reduce the amount of early enumeration in the Northwest Territories and to improve the quality of the data. To meet these objectives, we have to be concerned with several issues:

- People in the North want to be treated in the same way as people in Southern Canada. They do not think it's fair when the entire N.W.T. is enumerated by the long form census questionnaire (2B). They want to be counted through self-enumeration and sampling methods (one in five households) as are their southern counterparts.
- Native people are interested in obtaining questionnaires in their mother tongue, not just translation thereof. We will still have the problem of natives going "out on the land". There does not seem to be an alternative to early enumeration. Yet this in itself could be a problem. It is not clear legally if we can conduct an early census and require people to take part in it.

Development in technology, transportations and communications will continue to improve the enumeration of canvasser areas and also to enhance timeliness and data quality.

### Native People

The 491,460 individuals who identified themselves as native people in the 1981 Census made up just 2% of the total population. They are distributed very unevenly across Canada. Nearly 60% of the population of the N.W.T. is native, and about 20% of Yukon inhabitants identified themselves as native. Among the provinces, Manitoba and Saskatchewan had the highest proportion of native people - more than 6%. Alberta and British Columbia had 3% claiming native ancestry. (In the East, one person in 100 was identified as native.) Most native people live in the West.

Not all status Indians live on reserves. Indian reserves are territories set aside as a result of treaties between the Federal Government and the Indians. Fewer than 60% of status Indians reside on reserves. Four out of every 10 native people made their homes in urban areas compared with nearly eight out of 10 for other Canadians. There were significant differences among the native groups. Two out of 10 Inuit and three out of 10 status Indians lived in urban areas respectively compared to six out of 10 Métis and seven out of 10 non-status Indians. Their urbanization could however increase as a result of young natives migrating into urban areas in search of jobs.

The native population is a very young one, with an average age of 23 compared with 32 for non-natives. With the exception of the Inuit, native people use English as the language they speak most frequently at home. Educational levels among Canada's native people are lower than those of other Canadians, with only 25% of natives having at least high school diplomas as compared with 50% of other Canadians. Compared to other

Canadians, fewer native people are employed and/or in the labour force. They are more likely to have seasonal employment, and because of their relatively low level of education, they have been excluded from well-paying occupations. As educational levels continue to rise, native people will secure better jobs in the future, despite the distance from major labour markets.

Land claims and the drive for self-government could increase resistance to the Federal Government in general, and the census in particular. The native population generally claims that it has no input into enumeration procedures and that it obtains no direct benefits from the census, as per capita grants go to municipalities and not to the band. As municipalities generally do not provide reserves with any services, they do not see the benefits of it.

However, on the other side of the coin, if we are able to demonstrate the value of good statistical data and their importance, they may be more receptive to our data collection efforts.

**Table 1: Distribution of Native Population in 1981, Canada**

TOTAL NATIVE POPULATION	491,460
INUIT	25,390
STATUS INDIANS	292,700
NON-STATUS	75,110
MÉTIS	98,260

## **Résumé of the Question Period**

*After Mr. Kazmaier's presentation, a few issues, all of them related to telephone follow-up, were raised by the participants.*

### **Respondents' Concern about Privacy**

*First, there was a general concern about respondent sensitivity using the telephone for follow-up, especially in cases where the respondent did not give his phone number. This problem will worsen in the future as public concern about confidentiality increases in relation to the growth of personal information existing in computer form. Dependent as they are on public cooperation, censuses are vulnerable. Up to now, the censuses of United States and Canada have not been much affected by this potential problem.*

### **Impact of New Technology**

*Another concern relates to the impact of various technological developments on the use of telephones. The census already has to contend with the increasing use of answering machines. The United States Bureau of the Census (USBC) experimentation with centralized telephone follow-up has shown that the most useful way to handle these cases may be to leave a message and then continue to call. Unfortunately, experimentation in this area is still in its early stages.*

*A more positive technological development makes possible cross-checks from phone numbers to address geographic codes and then information. This development has great potential for the future. An immediate application of this capacity*

*would be to the Telephone Assistance Service during field collection, by enabling the operators to match an address to the geographic code when respondents call in to report that they have not yet received a questionnaire.*

### **Eventual Extension of Telephone Follow-up**

*The third concern was on the extension of telephone follow-up to resolve cases of failed-edit questionnaires and non-response households. The 1986 Test Censuses to be conducted by USBC will address this issue. The evaluation of test results will provide data on the coverage and data quality implications and also some idea of cost consequences.*

**Mr. Hicks'** presentation raised two major aspects of the extension of mail-back into areas currently enumerated using a pick-up methodology.

*One aspect is the potential impact of such a change on the Census of Agriculture. The extension of mail-back could have a deleterious effect on the coverage of farms in the Census of Agriculture because the pick-up methodology is largely used in the enumeration of farms, and enumerators play a role in ensuring good coverage.*

*The extension of mail-back into what are currently pick-up areas would also imply a redelineation of enumeration areas because an enumerator in a mail-back area can handle a larger number of households. However, the increased use of computers in generating geographic materials needed for collection could soften the impact of redelineation on the overall work-load.*



## **SESSION: CENSUS GEOGRAPHY**

Chairperson: D. Ross Bradley  
Geography Division  
Statistics Canada

Tuesday, October 8, 1985



# GEOGRAPHIC SUPPORT FOR THE 1990 DECENNIAL CENSUS

SILLA G. TOMASI

GEOGRAPHY DIVISION  
U.S. BUREAU OF THE CENSUS

## Introduction

The mission of the U.S. Bureau of the Census is to provide basic statistics about the people and economy of our nation to our Congress, our Executive Branch, and the general public. The success of a census rests not only on how well we collect data, but also on how well we link those data to geographic areas. Like pieces of an enormous jigsaw puzzle, the thousands of small geographic administrative units used to control census operations must fit together to include every square mile of the country, and each square mile must be covered once and only once. The millions of separate pieces of data then must be transformed into convenient statistical units that are arranged and rearranged in numerous combinations according to a variety of geographic areas needed to serve numerous interests. Without the ability to assign or relate data to specific areas, the data are of little value to the user community beyond that of national totals;

the data are interesting, but their usefulness is severely limited. Census data fulfill most user's needs only when related to a specific area, or a combination of areas, of the earth's surface; and it is the geographic framework of the census that makes the data meaningful by providing the identification of the geographic units, by name, boundary, and interrelationship (see Table 1).

## Geographic Support Function

It is not astonishing, therefore, to learn that the geographic support function is one of the most vital areas of concern in the planning effort for the 1990 Decennial Census. Geographic processes cut across, influence, or are influenced by almost all phases of census activity; they play a crucial role in every stage of planning, enumerating, and tabulating a census. Identification of geographic units and their boundaries is the basis for administrative control in taking the census and for processing and tabulating the results.

**Table 1. Types of Census Geographic Areas - 1980 Census**

POLITICAL AREAS	STATISTICAL AREAS
United States	Regions (4)
States & State Equivalents (57)	Divisions (9)
States (50)	Standard Consolidated
D.C. (1)	Statistical Areas - SCSA (17)
Outlying Areas (6)	Standard Metropolitan
Counties, Parishes, & Other	Statistical Areas - SMSA (323)
County Equivalents (3,231)	Urbanized Areas - UA (373)
Minor Civil Divisions - MCD (30,491)	Census County Divisions - CCD (5,512)
Incorporated Places (19,176)	Unorganized Territories (274)
American Indian Reservations (275)	Census Designated Places - CDP (3,733)
Indian Subreservation Areas (228)	Census Tracts (43,383)
Alaska Native Villages (209)	Block Numbering Areas - BNA (3,404)
Congressional Districts - CD (435)	Enumeration Districts - ED (102,235)
Election Precincts (36,361)	Block Groups - BG (156,163)
(in 23 participating States)	(Tabulated parts - 200,043)
School Districts (16,075)	Blocks (2,473,679)
Neighborhoods (28,381)	(Tabulated parts - 2,545,416)
ZIP Codes (= 37,000)	Traffic Analysis Zones (= 160,000)

Basically, all aspects of the geographic (and cartographic) support function evolve around two major tasks: first, the assignment of each residential address to its correct geographic location; and second, the classification of each location according to the various tabulation areas represented in the census. The mechanisms for accomplishing these goals entail provision of three major geographic tools: maps, address reference files – called GBF/DIME-Files in 1980 – and a geographic reference file – called the Master Reference File in 1980.

## Maps

Maps describe the earth in graphic form. Census maps show the streets, railroads, streams and other types of features a census enumerator would expect to see on the ground while collecting data for his/her assignment area. Maps also show the polygons – that we call blocks – formed by the features to which we assign unique numerical codes to represent the geographic entities of interest to the census. Once a census map has been marked with the field assignment area boundaries, a census enumerator can use this map to walk or drive around a block, enter every address that exists along each side of the block in a listing book, and record the number of the block in which that address is located. In that simple act, the enumerator has “geocoded” the address; that is, the enumerator has assigned it to a geographic location – the first of the two geographic support functions mentioned earlier.

## Address Reference Files

Whereas a map describes the earth in geographic form to a human, the address reference files (or GBF/DIME-Files that are similar to your Area Master Files) describe the earth in graphic form to a computer. Thus, the job once done by an enumerator now can be done by a computer. That is, the computer can perform the geocoding tasks for those areas of the country where this capability is feasible. In order to computer assign an address to its geographic location, it is necessary to capture the geographic information shown on a traditional census map – streets, railroads, streams, block identifiers, and so forth – and add information on the address ranges that apply to each side of a street between two intersecting map features. By adding the address range information, the computer can “see” what addresses fit into each block by using computer-matching algorithms that perform the geocoding task previously done by an enumerator. Because the computerized address reference files contain only geographic information and not information on the street length or position on the earth, a

map must be used when people participate in the geocoding process. For the 1980 census, about 55 million households, or about 60% of the total housing units in the census, were in areas covered by the address reference file geocoding capability.

## Geographic Reference File

Both of the geographic tools described so far have been concerned with the first of the geographic support functions – assigning an address to a geographic location. The second support function – classifying each geographic location according to the tabulation units recognized in the census – is performed by a geographic reference file, which we called the Master Reference File, or MRF, for the 1980 Decennial Census. This file shows, in computer-readable form, the relationship between and among the geographic units shown on the set of census maps covering the entire United States and its possessions. Using this file, we can relate the geographic location to which an address has been assigned – the specific block number written down by a field enumerator or assigned by the computer – to all higher-level geographic units for which we will tabulate data.

## 1980 Experience

All three of the geographic support products – maps, address reference files, and the geographic reference file – have several items in common; that is, they are simply three different ways of describing a part of the earth's surface. Problems with these geographic materials for the 1980 census caused confusion on the part of the Census Bureau's field staff and the data-using public. The problems resulted because all three of these geographic tools were prepared in separate, complex clerical operations using hundreds of clerks, many of them newly hired. Different errors were made on each product, leading to inconsistencies between the final products. One would expect mistakes to occur in an operation in which literally millions of geographic areas and identifiers were hand-entered on maps and in two independent, manually prepared computer files. As expected, we did not achieve perfect consistency among the three basic geographic products. The overwhelming majority of the identifiers in the three products matched perfectly, but where inconsistency did exist, it caused problems in data collection, tabulation of the data, and use of the data. Further, the labor-intensive processes by which the Census Bureau generated these geographic products contributed to their late delivery. The complex and functionally separate operations used to create the 1980 geographic products, like the processes used for earlier censuses, were not designed for the computer age.

In an effort to overcome the problems induced by manual and separate production of the geographic tools, the Geography Division has embarked on a program to produce a unified Geographic Support System. We have designed and are in the throes of building such a system for use in the 1990 census. This system will avoid many of the problems of the past. Our goal, therefore, is to record all available relevant geographic information about an area into a **single computer file**. This file will permit the computer to produce maps for field operations and publication, and permit the computer to perform the data tabulation operations for any geographic entity whose boundaries have been recorded in the file. We call this file, and its ancillary computer software, the Topologically Integrated Geographic Encoding and Referencing or "TIGER" System (see Figure 1). At the core of the system is the TIGER File.

An in-depth discussion of how we are building the TIGER System and an explanation of the TIGER File structure is beyond the purview of this paper. A brief synopsis, however, appears in order. If anyone is interested in further detail, please let me know, or write to Chief, Geography Division, Bureau of the Census, Washington, D.C. 20233. We are providing the Geocartographics Subdivision, Informatics Services and Development Division, Statistics Canada with copies of specifications for the TIGER File as they become available.

### The TIGER System

The fundamental concept of the TIGER File is the creation of a single geographic data base system with a computer-readable or digital map as its foundation. This file will cover the entire United States and the other areas for which the Census Bureau conducts a census. We created the internal structure of the TIGER File by applying the concepts of topology, a branch of mathematics that describes the spatial relationships of points, lines, and areas in a two-dimensional plane. Construction of the file using these principles allows us to fashion an elegant, powerfully self-checking computer data base that contains all basic map features and the boundaries, names, geographic codes, and address ranges of the various geographic entities. We store the geographic information in the TIGER File using lists and directories automatically linked and cross-referenced to each other. The line segments, points, and geographic area classification codes all are tied together so that when any one of them changes, the related components and relationships change, too (Marx, 1984). Integrated into this single digital map data base will

be all the information needed to allow (through the use of applications programs) census mapping by computer-driven plotting devices, the assignment of census questionnaires to geographic locations, and the cataloging and relational definition of areas. This means that all mapping and geoprocessing will be in complete agreement and that the problem with inconsistency among geographic products will be solved. The TIGER System also will allow us to produce maps of consistently high quality, which was not always possible in the past.

The TIGER File will be comprised of a series of records representing the position of roads, rivers, railroads, political and statistical boundaries, and other census-required features along with their attributes, such as name, geographic code, address ranges, feature class, and so forth (Broome, 1984).

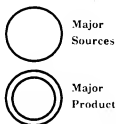
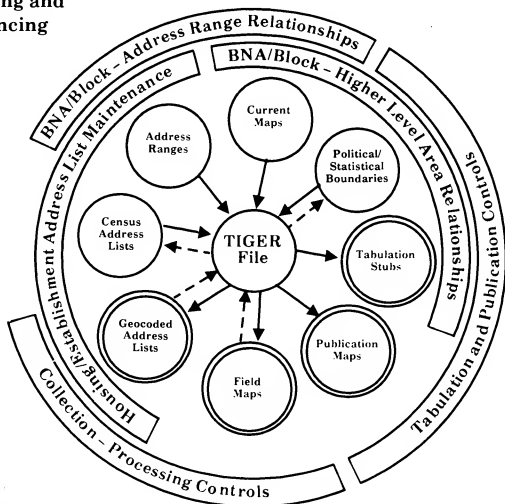
Each record in the file will contain information, such as:

- (1) the name and type of feature; that is, whether the feature is a road or street, a waterway, a railroad, a political boundary, and so forth;
- (2) the coordinate values defining intersection points along the feature, together with other geometric characteristics of the feature; for example, the curve vectors defining the shape of the feature;
- (3) the range of addresses located between intersection points for those records representing streets or roads, in addition to the post office name and ZIP code of each address range;
- (4) the codes of the geographic area(s) applicable to each segment of the feature based on the geometric relationship to the feature of the boundaries in the file for each geographic entity; and
- (5) other special situations associated with the record; for example, major employment centers or residential structures located along the feature.

Once constructed, the first, second and fourth characteristics of the TIGER File record provide a geographic framework from which the maps required for field operations or publication products could be generated; the first, third and fifth characteristics provide a means through which the addresses can be assigned to specific geographic areas; the first, third and fourth characteristics provide a basis upon which an

Figure 1. Components and Functions of the TIGER System

**T**opologically  
**I**ntegrated  
**G**eographic  
**E**ncoding and  
**R**eferencing  
System



automated questionnaire check-in and control system could be established, permitting generation of followup assignments based on structure address or serial number in geographic perspective; and the fourth characteristic provides the source file from which geographic table stubs and summary cartographic products could be generated for tabulation purposes.

Our charge is to complete the development of the TIGER File and the associated components of the TIGER System based on the 1990 Decennial Census schedule of activities; the first operation is slated for the first quarter of calendar 1988. We view the work to accomplish our objective in three broad stages, which can be viewed generally in the flow diagram shown in Figure 2.

- First, we must **create** the initial digital map data base – the underlying map image in computer-readable form – so that the computer will contain all of the streets, rivers, railroads, and so forth. We are working closely with our sister agency, the United States Geological Survey (USGS), to achieve this phase.
- Second, we must **update** this digital map data base with other information that makes the map useful for Census Bureau activities – adding new streets where subdivisions have been built; entering the names for all streets, rivers, and railroads so that Census Bureau field staff, as well as data users, can orient themselves; inserting the address ranges that go with each section of a street in the major urban areas; and adding the boundaries and codes of all the political and statistical areas for which the Census Bureau tabulates data.
- Finally, we must **use** the file and **enhance** this information based on the results of the early file-use operations. Here, the file-building and file-use activities interconnect because the file-use operations contribute valuable information from direct field staff observation to make the digital map data base match what currently exists on the ground.

Obviously, there are many other steps in the TIGER File building process and even more steps involved in using it to support the 1990 census. The ability to complete an ambitious job of this magnitude in such a short period of time will

involve expenditures of large sums of money and a commitment of literally hundreds of geographers, cartographers, computer programmers and clerks.

## Conclusion

The TIGER concept did not just erupt into existence. It came into being as a result of hard work and dedication on the part of a small core of people who, in 1981, were assigned by the Geography Division senior staff (Division Chief and Assistant Division Chiefs) to the full-time task of developing a long-range plan for the Census Bureau's geographic support activity. As expected, the major catalyst giving impetus to such an effort was the serious problem – still fresh in our minds – that existed in the delivery of cartographic and geographic products and services for the 1980 census.

The small planning group had at its disposal draft copies of two recently completed assessment reports of 1980 census operations prepared by separate interdivision work groups comprised of individuals who had worked on the census. It was apparent from the two reports that one major common denominator that threatened through the various operations was the need for and timely delivery of consistent geographic and cartographic support products.

The first result of our planning effort was a plan and budget completed in time for the Fiscal Year 1983 budget submission to the Census Bureau's Executive Staff. Without sufficient time to fully explain the ambitious plan, we were not successful in obtaining approval. Subsequent discussions and better documentation of the plan's benefits and budget considerations, coupled with more involvement of appropriate personnel from both the Department of Commerce and the Office of Management and Budget, culminated in approval of our long-range plan and the promise of enhanced funding levels beginning in the Fiscal Year 1984 (that is, October 1, 1983). Thus, in less than three years, we were successful in obtaining clearances to dramatically change how the geographic support function would be carried out for decades to come. We could only have achieved this goal through the existence of a long-range plan of action. Without the requisite front-end planning effort, there was little hope for launching and bringing such an ambitious geographic program to fruition.

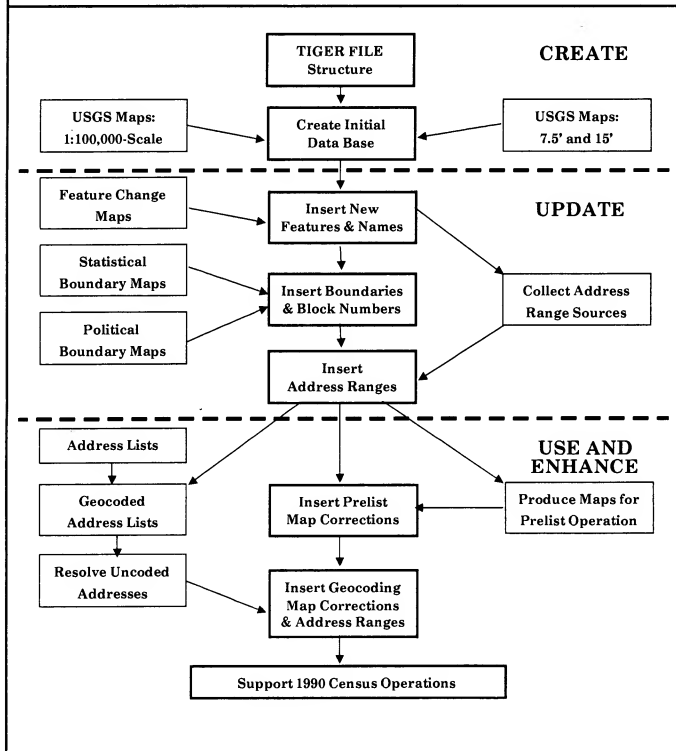
---

## REFERENCES

Broome, F.R., 1984. "TIGER Preliminary Design and Structure Overview, the Core of the Geographic Support System for 1990". 1984 Annual Meeting of the Association of American Geographers, Washington, D.C.

Marx, R.W., 1984. "Developing an Integrated Cartographic/Geographic Data Base for the United States Bureau of the Census". 1984 Ausra Carto One Meeting, Perth, Western Australia.

Figure 2. Stages to Create the TIGER File



# AREA MASTER FILES

## A Better Way to Serve Census Needs But How Far Should We Extend Coverage?

JOEL YAN AND KAROLE KIDD

INFORMATICS SERVICES AND DEVELOPMENT DIVISION  
STATISTICS CANADA

JEAN-PIERRE PARKER

GEOGRAPHY DIVISION  
STATISTICS CANADA

### 1. Introduction

To provide an understanding of the Area Master File (AMF) and how it has evolved since 1969, a brief description of the AMF concept is given together with some of the reasons behind its development. Secondly, the processes used for the creation and maintenance of an Area Master File are outlined. An overview of the evolution of the Area Master File program through the 1971, 1976, 1981, and 1986 Censuses is given which summarizes the major developments and achievements to date. Finally, some of the major issues surrounding the future of AMFs are discussed together with recommendations on how to deal with them.

### 2. Getting to Know An Area Master File

For over 15 years, the Geography Division of Statistics Canada has been creating and maintaining digital street network files known as Area Master Files. They form the geographic data base of a broader system, the Geographically Referenced Data Storage and Retrieval System (GRDSR), which has as its function the provision of geographic flexibility for census data retrieval.

#### 2.1 What Is the AMF?

As a geographic base file, the Area Master File is unique. An AMF is organized by census subdivision (CSD) and there may be one or more complete CSDs within one AMF. An AMF file comprises the digital definition of all city streets and other significant geographic features (such as highways, railroad tracks, rivers, hydro lines, etc.) which may be used in the delineation of special areas of interest.

Every AMF feature is defined in terms of three groups of information. The first group provides unique identification which takes the

form of census subdivision code, feature name (e.g., Albert St., Dow's Lake) and feature classification (e.g., street, lake, island). The second group defines the cartographic representation of each feature in the form of a series of nodes or points delineating, as accurately as possible, the shape and location of the feature as a single line image. Finally, the street (civic) address ranges for each portion of the feature (each block-face) are defined. A block-face is one side of a street between two consecutive feature intersections (see Figure 1).

In the file creation process, the block-face centroid is automatically calculated to represent the "block-face area" (all the addresses along a given block-face). It serves as the key through which census data for all the households on the block-face may be geographically referenced or "geocoded" for special retrievals by non-standard user-specified query areas such as school zones, police districts, etc. In urban areas, the use of block-face centroids has increased the geographic resolution for retrieval by approximately 25 times that for enumeration area (EA) centroids.

The block-face centroid also facilitates a high degree of historical continuity. By contrast, census enumeration areas change at such a high rate (approximately 40% per census) that historical continuity of data at this level of geography is not readily available. Other methods to provide more stability, such as use of a stable block as a geostatistical unit, are currently being analysed (Parenteau, 1985).

#### 2.2 Process Used to Create or Update the AMF

The process of creating an Area Master File starts with a set of accurate street maps obtained from a variety of sources but mainly

<b>CSD Code:</b>	<b>3544231</b>
<b>Classification:</b>	<b>Street</b>
<b>Name:</b>	<b>Stadium</b>
<b>Type:</b>	<b>Rd</b>
<b>Direction:</b>	<b>N</b>

Node Number: 012545  
Node UTM Coordinate:  
E-552124 N-4804067  
Intersecting  
Feature: Census Ave.

**Centroid UTM Coordinate:**  
**E-552112 N-4804041**  
 (Key to Census Data)  
**Address Range:** 375-495

from municipal agencies (see Figure 2). Large scale current maps showing block-face address ranges are required together with an up-to-date list of street names.

The manuscript maps are divided into sections which will fit on a digitizer, and a "node" is used to identify points wherever two features intersect and where features begin, end, or curve sharply. Each node is then identified by a unique serial number. In the next step, descriptive codes for every feature are transcribed onto a specially designed form. The information coded includes feature names, types, directions, node numbers, and addresses at intersections. The map manuscript is now ready to be placed on a digitizing table.

The digitizing equipment measures node positions relative to control points and generates one easting (X) and northing (Y) "table coordinate" for each node (see Figure 3). Since Universal Transverse Mercator (UTM) coordinates are used for AMFs, the table coordinates of the nodes must be transformed to UTM coordinates during subsequent computer processing. During the computer processing, the UTM coordinates for each block-face centroid are calculated using the coordinates of the nodes bordering the block-face (see Figure 4). Finally, all items and sections are merged to create an Area Master File for the census subdivision.

A series of error-handling and correction procedures are employed whenever Area Master Files are being created or updated. Extensive computer checking is done, to ensure that each node is linked to the correct feature segments and vice versa. This process locates the majority of clerical errors. When the file processing is complete, the file is plotted at the same scale as the original map. The two maps are then compared to ensure completeness and correctness of the AMF. Further plotting, followed usually by two or three update cycles, will produce a "clean" Area Master File. When input documents meet standards, the entire process of creating an AMF for a CSD of approximately 25,000 population requires approximately 40 person-days spread over 80 days elapsed time.

### 3. The Evolution of the AMF: 1969 - 1985

The evolution of the AMF from 1969 to the present time is described in terms of the ongoing development resulting in an increasing number of

products and applications. These developments are punctuated by status reports on the extent of AMF coverage at each census year.

#### 3.1 Conceptualization, Development and Initial AMF Creation

The conceptualization and software development phases for GRDSR and the AMF took place during the period 1969 to 1970 (Ion, 1969). A prototype AMF was built for the city of London, Ontario, in 1969 and was successfully used for testing.

The longer term objective was complete AMF coverage for census subdivisions within all census metropolitan areas (CMAs) and census agglomerations (CAs) with over 50,000 population. However, in the initial stage of AMF creation from 1970 to 1971, priority was given to the creation of parts of the 14 main CMAs and CAs based on demands from municipal governments who were prepared to collaborate with Statistics Canada. As a result, the geocoding methodology and some of the software were transferred to, and applied by, several major municipalities in building their own geographic information systems (Richmond, 1974).

#### 3.2 The 1971 Census

During the 1971 Census processing cycle, the data for households on each block-face were successfully geocoded to the block-face centroids of the existing AMFs. As demand increased, AMFs for municipalities surrounding the cities of Vancouver and Winnipeg were created in 1973, and 1971 Census data were retroactively geocoded to these centres. In all, block-face geocoding covered a total of 54 CSDs and approximately 7.5 million or 34% of the Canadian population for the 1971 Census (see Table 1). These 54 CSDs were in effect subdivided into approximately 200,000 micro areas which brought much more flexibility to the operation of special area retrieval or aggregation than was possible with the EA (Facts by Small Areas, 1972).

#### 3.3 Major Expansion in AMF Coverage

In preparation for the 1976 Census, a program was undertaken to continue the extension of block-face geocoding in collaboration with municipalities. Because of this collaboration by the municipalities in supplying documents and performing encoding activities while

**Figure 2. Input Base Map**

The map illustrates the input base map with the following features:

- Water Bodies:** Moustafa River, Wilson River, King Lake, and a Stream.
- Infrastructure:** Energy Line, Highway (HY), Macdonald, Stadium, Cherry Rd, Victoria Rd, Harvest, and a Dam.
- Landmarks:** Philemon University, Park Limit, and a Green area.
- Other Labels:** CN, RD N, and a dashed line indicating a boundary or limit.

Figure 3. AMF Output: Single Line Image

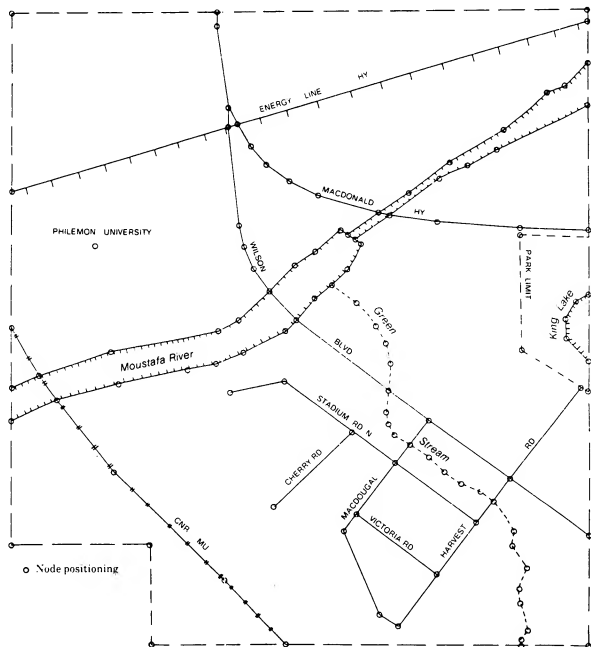
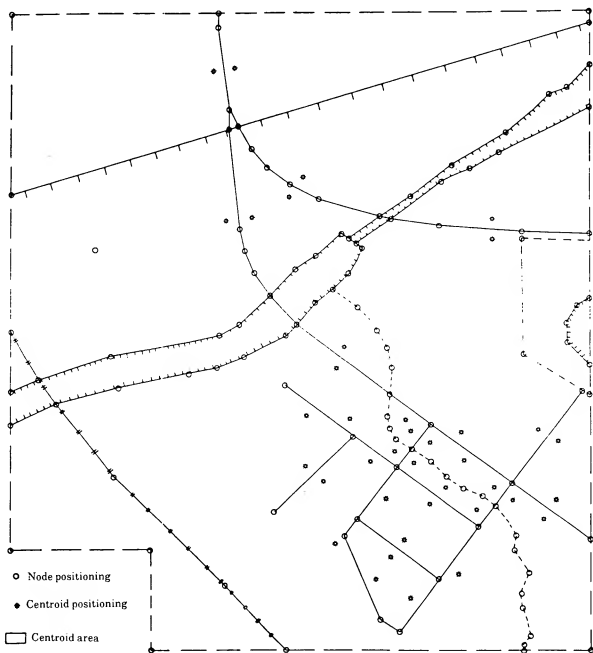


Figure 4. Block-face Centroid



**Table 1. Area Master File Coverage by Census (1971, 1976, 1981)****(Number of Census Subdivisions, Population Covered and Percentage)**

Province	1971 Census			1976 Census			1981 Census		
	CSD	Pop.	%	CSD	Pop.	%	CSD	Pop.	%
Nfld.	-	-	-	1	86,576	15.5	1	83,770	14.7
P.E.I.	-	-	-	-	-	-	-	-	-
N.S.	1	122,035	15.5	2	183,223	22.1	2	176,871	20.9
N.B.	-	-	-	1	85,956	12.7	2	135,264	19.4
Que.	12	1,677,611	27.8	58	3,277,026	52.6	59	3,233,410	50.2
Ont.	12	3,055,090	39.7	40	5,082,079	61.5	48	5,496,971	63.7
Man.	12	535,217	54.2	1	560,874	54.9	1	564,473	55.0
Sask.	2	265,918	28.7	2	283,343	30.8	2	316,823	32.7
Alta.	2	841,471	51.7	2	931,278	50.7	2	1,124,989	50.3
B.C.	13	877,956	40.2	18	1,174,432	47.6	42	1,580,190	57.6
N.W.T.	-	-	-	-	-	-	-	-	-
Yukon	-	-	-	-	-	-	-	-	-
Canada	54	7,375,298	34.2	125	11,664,787	50.7	159	12,712,761	54.5

Statistics Canada performed map digitizing, central processing and assembly of the information into AMFs, it was possible for more AMFs to be created and maintained for the same relative level of funding.

### 3.4 The 1976 Census

The resulting increase in block-face geocoding coverage was on the order of 4.3 million or 16% of the population, reaching a total for Canada of approximately 11.6 million or 50.7% of the population. The number of census subdivisions represented increased to 125 (see Table 1), and they were subdivided into approximately 350,000 micro or block-face areas.

### 3.5 New AMF Products and Applications Are Demonstrated

With the original mandate of servicing non-standard areas of interest adequately fulfilled, the development of new services which would make use of the specialized information contained in the AMF was initiated. As well, AMF coverage began to be extended, based on demand, to centres smaller than 50,000 population.

#### Automated Street Indexes

The creation of street indexes, which had previously been a time-consuming and an error-prone manual activity, was automated for the Chief Electoral Office and census field

operations. In the case of the Chief Electoral Office, it was possible to quickly identify the federal electoral district to which an elector should go simply by looking up the elector's address in a custom-made street index. Similarly, in census field operations it was very easy to identify the appropriate Census Commissioner and Enumerator for follow-up or when a householder reported that he/she had not received a census questionnaire at a specific address.

### Prototype Computer-assisted Collection Mapping System

The conceptualization and development of a prototype system to automatically generate double line street maps from the single line image of the AMF permitted the automated generation of collection maps for approximately 34 CTs (Yan, Bradley, 1982). While this experience did not provide immediate production cost-savings, it did demonstrate the feasibility of the task and the benefits possible with further system improvements.

### Applications in Other Agencies

Other agencies, including the Stratford police department and the city of St. Catharines Engineering Department for road inventory (Loewen, 1979), began using AMF as their basic cartographic file.

### 3.6 The 1981 Census

For the 1981 Census, AMF coverage had been extended to 159 CSDs and approximately 12.7 million or 54% of the total Canadian population was covered through the block-face geocoding program (see Table 1). Small area data retrieval was now possible for approximately 18,000 EAs or approximately 500,000 micro units with households attached to them.

### 3.7 Recent Initiatives

More and more concrete benefits to census and other Statistics Canada operations are emerging through applications aimed at exploiting the AMF. Cooperation with federal and provincial mapping agencies is growing, and joint ventures with municipal governments and other agencies to extend AMF coverage are on the increase. The following are brief descriptions of some of these initiatives.

#### Collection Mapping for the 1986 Census

In 1983, an AMF collection mapping production system was developed based on the initial prototype system described in Section 3.5. For the 1986 Census, 1,200 census tract collection maps representing approximately 6,000 EAs have been automatically produced at a cost of \$60 per map compared to a manual redrafting cost of \$200 per map.

As a secondary product, this system permits the production of double line maps with such data as number of households and postal codes plotted at the block-face centroid.

#### Postal Code Conversion File

Software has been designed and successfully implemented to link the postal code file containing 6-digit individual postal codes to the corresponding block-face centroids of an AMF file using the address ranges. With this facility it is possible for census data to be extracted by areas defined by postal codes.

A postal code conversion file (cross-referencing postal codes to 1981 standard geostatistical areas) has also been developed, thereby permitting the extraction of data from administrative files for selected standard geostatistical areas.

#### Block-face Data Linkage File

A special block-face file showing household and population counts has been provided to

Labour Force Survey for use in the re-design of their sample units.

#### Joint Ventures to Increase Coverage

Since the 1981 Census, there has been an increased interest shown by external clients in digital street files. This interest has led to several joint ventures aimed at extending AMF coverage (Yan and Parker, 1985).

One venture has resulted in major extensions to AMF coverage in southern Ontario and allows its use as the basic geographic file and a unique cartographic base for emergency agencies to make possible the computer-aided dispatch of emergency vehicles. Most of the eight person-years effort involved was provided by police departments.

Several joint ventures have been initiated which involve group efforts in order to reduce duplication and share the actual work. One such venture with Metro Toronto will see them updating their digital network file to reflect 1985 status. The completed file will then be used to update and/or replace Statistics Canada's 1981 version. Other similar ventures are being negotiated with municipalities such as Burnaby, Winnipeg (Courage et al., 1982) and Calgary (McNabb, 1982).

A memorandum of understanding between the Ontario Ministry of Natural Resources, the county of Oxford and the city of Woodstock, and the Geography Division of Statistics Canada has been drawn up for creation of an AMF for the city of Woodstock.

#### Pilot Project With EMR

A pilot project was undertaken with Energy, Mines and Resources (EMR). It involved the loading of an EMR digital file containing the UTM coordinate values for features, and then adding the attributes necessary to create the AMF (Yan et al., 1985).

#### Increased Cost-recovery Work

The sale of AMFs through a licensing agreement is now being negotiated as a means of offsetting creation and maintenance costs.

The collection mapping system has been enhanced and used to generate more than \$50,000 in cost-recovery for mapping and street index work done for ambulance dispatch in Ontario.

Thus, it can be seen that the scope of applications is broadening to the point where existing production mechanisms, file structures and systems capacities are reaching their limit. It is appropriate at this time to step back and review the issues related to future directions for the AMF.

#### 4. The Future of AMF

In 1969, when the AMF was first designed as the geographic nucleus of GRDSR, the extent to which AMF would evolve as a stand-alone product or a link through other files to such a wide range of applications may not have been foreseen. It is time now in the mid-1980s to re-assess AMF and plan for its future development and expansion as we move toward the 1991 Census and beyond. The issue is not only how far to extend AMF coverage, but also how to manage that extension. Some alternatives and recommendations are presented in this section.

##### 4.1 The Need for AMF Extension

The recent trends and experience with specific applications in the census context and beyond all point to the need for continued extension of Area Master File coverage. The following examples illustrate this point.

1. The AMF is now the base not only for census data retrieval but also for census cartography, street indexes and other products. Where computer-assisted collection mapping (CACM) has been implemented, significant savings have been realized (costs reduced to \$60 per map compared with \$200 per map for manual re-drafting). In addition, the collection mapping base is compatible with the base used for census retrieval, thereby resulting in a reduction of person-years needed for manual reconciliation and the maintenance of two bases. The existing AMF coverage limits the extent to which these kinds of savings are possible.
2. Users requesting non-standard area data retrievals have much greater flexibility through AMF block-face coverage than they do in urban areas without AMF coverage. As a federal agency, we should keep in mind the potential political repercussions or implications that these two levels of coverage (and therefore service) provide.

3. The provision of census data by postal code appears to be a very lucrative market; however, to date the full potential of the postal code link to AMF has not been exploited. The postal code may very well be the flexible type of geographic locator needed to provide the capacity to update data to current geographic limits of cities, provided the postal code conversion file is maintained. However, this is dependent on the existence of an AMF. Hence, expanded AMF coverage would make the use of the postal code more attractive (Puderer, 1985).
4. The AMF is being proposed as a framework for attaching individual addresses that would form part of an address register (Royce, 1985).
5. Specific external users such as police departments and ambulance services are using the AMF for computer-assisted dispatch (CAD) and are cooperating with Statistics Canada to extend AMF files. New demands by these agencies for extensions are coming in regularly.

##### 4.2 Scenarios for Managing the Extension of AMF Coverage

A number of important issues must be considered within the context of planning and managing the extension of AMF coverage. They are as follows:

1. the scope of AMF applications;
2. Statistics Canada's role in producing AMFs and coordinating the production of AMFs by others;
3. the logical and practical limits to which the extension of AMF should be attempted;
4. the content of the AMF in relation to Statistics Canada's needs and the structure of cartographic data bases produced by other agencies; and
5. the degree of software modification.

For each of these issues, various alternative scenarios and recommendations are put forward. It is hoped that by discussing these issues now, effective decisions can be made regarding the future of AMF for the 1991 Census and beyond.

## Issue 1: The Scope of AMF Applications

### Alternative scenarios:

- 1.1 AMF for census applications only;
- 1.2 AMF for all Statistics Canada's applications;
- 1.3 AMF for Statistics Canada and outside agencies on demand (especially emergency agencies, municipal agencies).

### Recommendations:

1. Extend AMF applications for census based on AMF as the central urban geographic framework. New applications might include automatic geocoding of place-of-work data, census block program, and land-based Census of Agriculture.
2. Promote and demonstrate AMF as a key geographic tool for other operations within Statistics Canada, e.g., address register, small area data program, administrative data program, postal code linkage.
3. Develop effective external applications on a cost-recovery and "good public servant" basis.

## Issue 2: Statistics Canada's Role in AMF Extension

### Alternative scenarios:

- 2.1 only census personnel involved in AMF production;
- 2.2 only Statistics Canada personnel involved in AMF production;
- 2.3 census and selected external parties who contact Statistics Canada involved in AMF production;
- 2.4 conduct an organized national/provincial program to solicit partners in AMF extension.

### Recommendations:

1. Census should ensure that a minimum standard of extension is maintained for its own needs through its own resources.
2. Census should encourage extension by agencies who have better local knowledge wherever there is a local interest

and adequate technical capability. Strong local support is recommended.

3. Statistics Canada should encourage participation through developing:
  - (a) improved documentation packages for local creation and support of AMF;
  - (b) other tools and basic application software to increase utility to other agencies.
4. Statistics Canada should serve the central role of:
  - (a) setting minimum standards for file content and exchange;
  - (b) organizing the work and the communications;
  - (c) serving as clearing house for software and information exchange.

## Issue 3: Limits to AMF Coverage

### Alternative scenarios:

- 3.1 no change in AMF coverage;
- 3.2 extend to a consistent geographic level based on resources available, e.g., to complete CMAs or CAs, or to fixed population of CSDs, or to the level of door-to-door postal delivery;
- 3.3 extend based on demand and support - only external market;
- 3.4 extend based on a formula which considers demand and utility;
- 3.5 extend to national level over time based on cooperative ventures with major partners (à la TIGER).

### Recommendations:

1. Geography should consult with census and other programs within Statistics Canada to determine the utility of extending AMFs to different regions of the country - priority 1, priority 2, priority 3, etc.
2. Develop a formula for the utility of AMF extension based on the resulting benefit to Statistics Canada and the cost to the department if local or other agencies are willing to share the cost or create a fund.

3. Identify key agencies who are undertaking or who plan to undertake major digital mapping programs. Select those agencies for whom the content is similar and the accuracy levels are acceptable to Statistics Canada. Move through joint programs towards a national consensus. Develop interfaces and test the feasibility as soon as possible. Plan major extensions for the 1991 Census. Involve top management.

#### **Issue 4: Degree of Change in AMF Content**

##### **Alternative scenarios:**

- 4.1 freeze at current content;
- 4.2 undertake minor enhancements to content as the need arises;
- 4.3 be prepared to extend the content in line with the needs of major users and partners.

##### **Recommendations:**

1. Maintain the current basic concepts of the network file and maintain an interface to the current AMF format, no matter what content enhancements are made.
2. Continue the process of gradually upgrading the AMF content to permit an improvement to current products, e.g., major roads and major buildings.
3. Before the 1991 Census, undertake a major review of the geography infrastructure to improve the structuring between levels (network, boundaries, postal code) in the infrastructure, and streamline maintenance and handling. A detailed look at the TIGER system should be included.
4. If and when partners for major extensions are identified, be prepared to discuss enhancements to AMF content in proportion to the degree of return services offered by the other agency. However, maintain the central concept of the network file as the primary base with dependent but separate layers for attributes, postal codes, boundaries, blocks, etc.

#### **Issue 5: Use of New Technology in AMF Maintenance Program**

##### **Alternative scenarios:**

- 5.1 be reactive in system maintenance, i.e. maintain current systems except for correction of bugs;
- 5.2 be proactive in system development.

##### **Recommendations:**

1. Streamline methodology/techniques to take advantage of recent technological developments and therefore be able to do more with existing resources.

#### **5. Summary of Recommendations**

The previous section outlined in some detail the range of issues surrounding the continued development of AMFs. The key recommendations are summarized here.

1. Maintain present support of AMF program.

Given the successful extended role of the AMFs within census geography and other projects, ensure ongoing funding to at least maintain the basic AMF program.

2. Continue development of AMF applications.
3. Extend AMF coverage through ongoing Statistics Canada resources and joint ventures with major partners.
4. Review and enhance AMF content in line with requirements of major users and partners to facilitate their utilization of the files.
5. Develop a long-term plan for AMF use.

Given the changing role and the demonstrated benefits that can be generated through the use of the AMFs, a long-term strategy for AMF development should be undertaken. Major components of this plan should include:

- (a) analysis of new application of AMFs within census, Statistics Canada and outside agencies;
- (b) an exhaustive survey of other agencies in the digital mapping business;
- (c) formulating policies to facilitate and support "joint ventures".

6. Provide additional budget for AMF Research and Development.

Research and development are required to:

- (a) improve infrastructure;

- (b) expand the product line;

- (c) streamline systems and methods; and

- (d) facilitate joint projects.

7. Develop a promotional and marketing plan.

---

#### REFERENCES

- Boisvenue, A. and R. Parenteau, 19--. **The Geocoding System in Canada and its Area Master File**, pp. 226-232.
- Courage, W.G. and P. Bennett, 1982. "A Practical Information System Developed With Limited Funds", **papers from the Annual URISA Conference**, pp. 159-172.
- Geography Division, Statistics Canada, 1983. "Block-face Geocoding Coverage for the 1971, 1976 and 1981 Censuses", Working Document No. 6.
- Ion, R.J., 1969. "The Geographic Basis of the DBS Geocoding System for Urban Areas: An Overview", Analytical and Technical Memorandum No. 3, Dominion Bureau of Statistics, Ottawa.
- McNabb, G.H., 1982. "Spatial Data, Geographic Referencing and Computer-assisted Mapping at the City of Calgary", **papers from the Annual URISA Conference**, pp. 233-241.
- Parenteau, Robert F., 1986. "A Block Program - Yes or No?", **Proceedings of the International 1991 Census Planning Conference**, Statistics Canada.
- Puderer, Henry A., 1982. "Census Geography Staff's Planned Postal Code Master File - Area Master File Linkage", paper prepared for Canadian Geocoding System Workshop as part of the Canadian Association of Geographers sessions during the 1982 Learned Societies Conference, University of Ottawa.
- Puderer, Henry A., 1984. "Postal Codes and the 1981 Canadian Census of Housing and Population", presentation to the Urban and Regional Information Systems Association Meeting held in Seattle, Washington, August 12-15, 1984. Revised: 84.8.30.
- Puderer, Henry A., 1985. "Statistics Canada's Postal Code Products and Services", presentation at the 23rd Annual Conference of the Urban and Regional Information Systems Association held in Ottawa, Ontario, July 28 to August 1, 1985.
- Richmond, D.E., 1974. "The Joy of George - A Municipal Guide to Geocoding", Management Systems Development Department, City of Calgary, 16 pages.
- Royce, Don, 1986. "Applications of an Address Register in the Canadian Census", **Proceedings of the International 1991 Census Planning Conference**, Statistics Canada.
- Statistics Canada, 1972. **GRDSR: Facts by Small Areas**.
- Yan, J.Z., and D.R. Bradley, 1982. "Computer-assisted Cartography for Census Collection: Canadian Achievements and Challenges", **Proceedings of the Sixth International Symposium on Automated Cartography**, Washington, March 1985, pp. 584-599.
- Yan, Joel Z., and Jean-Pierre Parker, 1985. "A Framework for Coordinating the Development and Application of Street Network Files for Canada", **papers from the Annual URISA Conference**, pp. 132-143.

## A BLOCK PROGRAM

Yes or No?

ROBERT PARENTEAU

GEOGRAPHY DIVISION  
STATISTICS CANADA

### Introduction

Development of the geographical structure for the conducting of a census of population and housing is a very complex operation. For the purpose of collecting information from all Canadian households, the entire country is subdivided into small units called "enumeration areas" (EAs). The EAs, which are the geographical units enumerated by Census Representatives, **change with each census.**

The instability of the EAs from one census to another is the focus of this presentation. The fact that the enumeration areas are constantly changing is a source of ineffectiveness for the following reasons:

1. difficulty in conducting longitudinal studies, since the summary tapes are prepared by EA and the "reverse record check" evaluation methodology for tracking households in the previous census is based on EAs;
2. need to completely recode the geographical files (particularly the Census Geographic Master File), which are based on EAs;
3. reduced flexibility in the description of the territories defined by statistics users;

4. strong impact on other activities at Statistics Canada (for example, the Labour Force Survey);
5. creation of operational constraints which go against the common desire to decentralize, at the regional level, delineation of enumeration areas;
6. inability to control changes (modification of EA delineation criteria by regional operations, revision of geostatistical units and adjustment of EA boundaries at the request of users); and
7. less-than-optimal delineation of certain geostatistical units (CTs, PCTs and urban areas, for example) because the collection operations call for maximizing the EAs, and this results in changes in the boundaries of the geostatistical units for operational reasons (see Table 1).

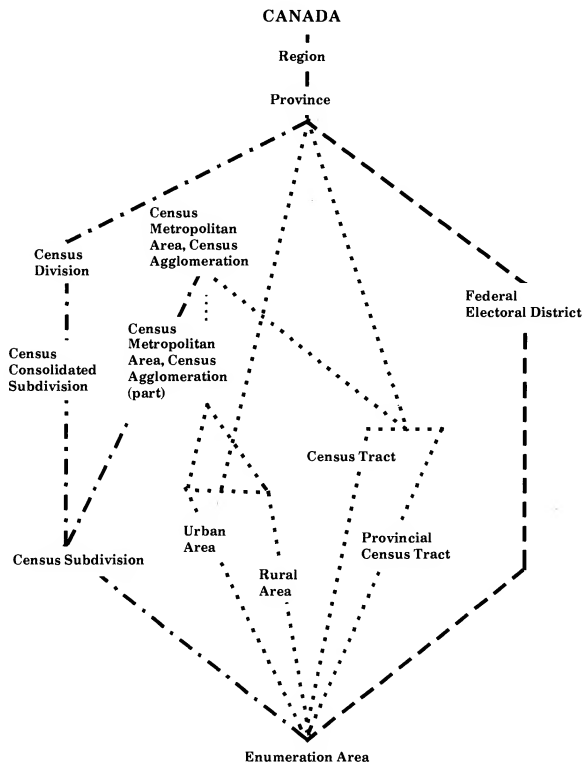
### 1. Past Experience

In order to understand better the importance of the enumeration area, the reader should examine Figure 1, which shows how the EA is situated with respect to the geostatistical units and indicates the vital role it plays in data compilation.

Table 1. Number of EAs\* (1986 Methodology)

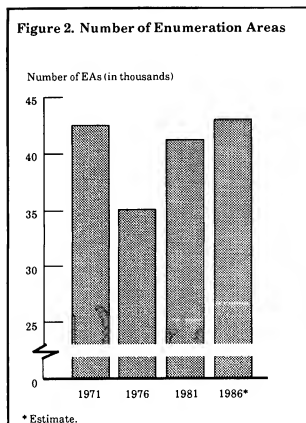
Methodology	>		<		Average	Total
Mail-back (criterion: 375 dwell.)	2,511	12%	18,851	88%	284.70	21,362
Pick-up - Urban (criterion: 300 dwell.)	658	12%	4,762	88%	160.06	5,420
Pick-up - Rural (criterion: 175 dwell.)	3,068	20%	12,185	80%	109.25	15,253
Total	6,237		35,798			42,035
*85-09-30						

Figure 1. Census Geostatistical Area Hierarchy



Delineation of EAs is based on federal electoral districts (FEDs). A FED is a territorial division from which a member of the House of Commons is elected. The FEDs are modified after each decennial census to reflect changes in population distribution; their boundaries may extend beyond those of census divisions and subdivisions, but not those of provinces. The 1991 Census should be based on a new representation order.

Since an EA must respect the boundaries of the various geostatistical units, which may be changed several times during the intercensal period, it is very unstable. For the last three censuses, the number of EAs fluctuated widely: there were 42,533 in 1971, 35,154 in 1976, and 41,197 in 1981. It is expected that there will be approximately 43,000 for the 1986 Census (see Figure 2).



This variation is largely the result of changes in EA delineation criteria. The EA network is not stable enough to ensure the obtaining, through a given census, of information regarding a specific territory with the boundaries it had in the previous

censuses, since those boundaries change. The EAs of a given census cannot, in general, be superimposed on those of any preceding census. Nor does their delineation respect the boundaries set for the geostatistical units in the previous census or censuses, if those boundaries have changed. The reason for the instability of the EAs is the fact that they are used in more than one way: (1) for the purpose of field data collection, they are administrative and operational units, (2) for the purpose of data dissemination, they are statistical units. Each function has a certain number of specific basic requirements, and these requirements are sometimes divergent.

In order to solve this instability problem, it was decided that, for the 1986 Census, the boundaries of the rural as well as urban enumeration areas would be the same as those for the 1981 Census, wherever possible. The aim of this policy was to make the EA boundaries for the 1986 Census stable with respect to those for the 1981 Census. We lost many of our illusions when preparing for the 1986 Census (see Figure 3). Enumeration areas, by virtue of their many functions, are not stable units: only 62% of the 1986 EAs have boundaries comparable to those of the 1981 EAs.

## II. Alternatives to Enumeration Areas

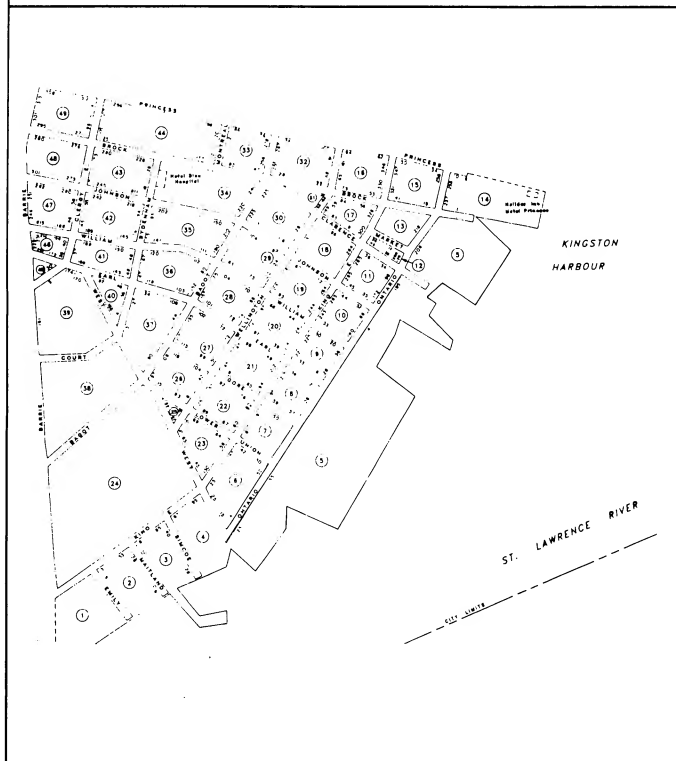
Enumeration areas in their present form do not meet the stability and flexibility requirements of the data-dissemination function.

Removal of the dissemination function would help facilitate the creation and delineation of the enumeration areas. The task, therefore, is to find an alternative. All of the possible solutions depend on various factors and contingencies - for example, the time, staff and budget available, the possibility of using existing elements, methods or equipment, and the possibility of reusing the units from a previous census (the 1986 EAs, for instance) as a point of departure for a new system allowing for data comparison beginning with this census.

I will describe briefly a few of the various solutions that might be considered.

- (i) Use of a regular grid: a regular grid would generate permanent geographic areas that could be used as elements for the reconstitution of dissemination entities in the development of statistics.

Figure 3. Computer-Assisted CT Map



- (ii) Use of the EAs of the previous censuses as the basic units: the EAs of the previous censuses would be geocoded and would become the data-dissemination base for future censuses. We would have a network of permanent units which could be superimposed on the new EAs at the time of each census and serve as a basis for chronological comparison. These basic units would allow for reconstitution of the dissemination entities.
- (iii) Use of postal codes: Great Britain is currently studying the possibility of using postal codes as data collection units. Since the postal system covers the entire country and the postal codes are linked to the dissemination units, postal codes could become the operational unit for collection.
- (iv) Delineation of blocks: the U.S. Bureau of the Census plans to divide the United States into blocks for the 1990 census. This objective is an extension of programs established during the last two censuses (1970 and 1980). The Canadian Census already uses the block concept, although a clear definition has not yet been developed.

The first solution would allow for delineation of stable, flexible and permanent units. However, the establishment of such a system would certainly be a costly undertaking. This new technique would create a considerable break in the development of the Canadian geographical structure. I do not think that we are ready for this. In addition, integration of a regular grid with the irregular boundaries of the blocks forming the geostatistical units would cause a tremendous number of problems.

The second solution would be much less revolutionary. The principle of construction by predetermined element already exists for enumeration areas. The defined unit would be stable and flexible and could be applied throughout Canada.

The third solution, which involves using postal codes as collection units, could be applied where there are Area Master Files - that is, in Canada's major urban centres. "Postal code and block-face" matching has been done in those centres. However, this method could not be applied in the small urban centres without enormous investments to create other AMFs, and use of it

would be out of the question in rural Canada, where postal codes cover a large territory. This solution should not be rejected completely; a step-by-step application of it has potential benefits.

The last solution, that was chosen by the U.S. Bureau of the Census, involves dividing the territory into blocks. An examination of this option in light of the American experience follows.

The solution or combination of solutions chosen must ensure historical continuity as much as possible, in order that a sharp break in statistical chronology may be avoided. Such a break would likely lead to problems in the establishment of links in space and time. For example, it would likely be difficult to satisfy the users' valid need for linkage of their statistical series based on EAs (summary tapes based on EAs have been announced and sold since the 1971 Census).

Delineation of blocks might be the most effective solution. A first city-block conceptualization and delineation draft was developed for the 1971, 1976, 1981 and 1986 Censuses.

### III. Experience With Blocks in the Censuses from 1971 to 1986

Definition of blocks is not really a new idea. The Geography Division considered the question in 1971 and 1976, when the census block was presented as a tool defined by Statistics Canada to make census information available for the smallest areas possible, while respecting the restrictions set out in the Statistics Act. For the 1971 Census, a block was defined as being the smallest area surrounded by streets. A block was generally rectangular, but it could also have an irregular shape and have boundaries of another kind - for example, a waterway or a political boundary. For the 1976 Census, the definition became more precise: a census block was the smallest enclosed area that could be delineated within a census tract or provincial census tract, and census blocks could vary from one another in terms of population, boundaries, shape and size (see Figure 3).

An assessment of the blocks for the 1971 and 1976 Censuses has already been carried out. The following is a summary of the most important points noted in this assessment.

- (i) Use of block numbers was connected mainly with data collection and processing. The enumerators were to proceed block by block; the block numbers were entered in the Visitation Record, facilitating the geographical activities carried out immediately after Census Day.
- (ii) Use of block numbers for dissemination was considered a secondary objective.

In light of the 1971 and 1976 experiences, it was recommended that, for the 1981 Census, only census subdivisions that were included in the census tract program be subdivided into blocks. This recommendation was observed, but the block was never defined as a geostatistical unit. The fact that the 1981 Census Dictionary contains no definition of the term underlines the lack of importance assigned to the concept. Some census data (population and number of households) are available by block, but there has been no publicity or marketing in that regard. The budgetary restrictions and cut-backs affecting the Geography Division in the late 1970s delayed the development of any new geostatistical concept.

In the summer of 1982, it was decided that, for more than 1,200 census tracts included in the Area Master Files (AMFs), delineation of blocks (and EAs) would be a computer-assisted operation based on the features in the AMFs. AMFs are machine-readable documents which describe the layout of streets and other physical features using a system of co-ordinates.

For the computer-generated collection maps of the 1986 Census, blocks are therefore determined on the basis of all the linear features in the AMFs, excluding property lines, marshalling yards and railway sidings. The block numbers are placed between these features, in the positions defined by the centroids. The numbers are assigned to the individual blocks in ascending order - that is, 1, 2, 3 and so on. If a block is too small, the number is not automatically drawn. The block numbers are therefore not in perfect sequence; block numbers are also assigned to lakes, rivers and islands. The numbers are assigned from west to east and from south to north.

With computerized production of census tract maps and delineation/numbering of blocks, the cost of preparing census-tract

maps can be greatly reduced. The estimated cost of producing CT maps by computer is \$60 a map, compared with \$200 for manually produced maps. Barring any unforeseen circumstances, computer-assisted collection mapping will be used even more for the 1991 Census. All of the census tracts (approximately 40,000) should be produced automatically, based on the AMFs. Given the continued efforts to improve the contents of the AMFs and to broaden their application, computerized delineation of blocks should cover a large proportion of the Canadian population in 1991 (approximately 65%).

#### IV. Strategy for Future Censuses

Large amounts of time and money have been invested in the development of blocks for recent censuses, but these investments have not always been productive for lack of a well-defined policy.

The question that must be asked at this conference is whether Statistics Canada should finance further projects to determine the advisability of a block program. In the U.S., blocks have been defined and delineated since 1970 as complements to enumeration districts. For the 1990 census, the policy adopted by the U.S. Bureau of the Census is that the block program will be extended to cover the entire United States. Blocks and enumeration districts will be defined either as dissemination units or as administrative units, and the two types of units will be mutually exclusive. All areas for which data are available by block or by group of blocks will have enumeration districts as administrative units, and vice versa.

This last point is very important. It is the answer to one of the problems connected with enumeration areas - namely, the instability resulting from the dual role assigned to the EAs. If dissemination units and collection/administrative units are defined independently, it is easier to control the changes affecting each type of unit.

American census specialists concluded from the 1980 census experience that control of collection activities at the block level allows for better supervision of the entire task and for improved national coverage. Availability of data for the entire country at the block level gives users the flexibility of aggregating data for the territory in which they are interested.

This notion of data-aggregation flexibility for users is a central issue. Do the users want to obtain census data by block or do they want to maintain the status quo? For a user accustomed to obtaining statistics by enumeration area, introduction of a new dissemination unit could be a major source of problems and confusion.

Consultation with the users is therefore advisable. It could initially be done at the regional office level, where there is regular contact with the users of the statistics. If increased use of the statistics is desired, the users must believe that the information could be of assistance in their day-to-day decision-making.

In order to prevent negative reactions on the part of local users, the U.S. Bureau of the Census encouraged local organizations to submit proposals regarding the delineation of enumeration districts and blocks. The Canadian census tract program operates on the same principle. These same contacts could become the co-ordinators for a block program in the major urban centres.

This last issue is really the most important one: a block program will be developed only if it meets the needs expressed by the users of census data. If the users prefer the status quo, a special unit could be defined solely for use in data collection. It would not become a dissemination unit.

If the users believe that the block program would be of great benefit to Canadians, Statistics Canada should develop a national

policy. For the 1991 Census, there would be a pilot project in which all of the relevant questions would be examined: definition of the term "block", boundaries, numbering, matching with the EAs of previous censuses, dissemination of data by block or by group of blocks, and confidentiality of statistics. In addition, Statistics Canada should look at what has been learned in this area by the Americans. Intensive discussions with census specialists in the United States are called for.

Other questions must also be considered. What are the advantages of the establishment of such a program for Statistics Canada as a whole? The **Labour Force Survey** could benefit enormously from permanent numbering of blocks, which are the basic sampling unit. Sampling is now based on the blocks (and block numbers) of the 1981 Census. In order to update the figures for the number of households per block to reflect the results of the 1986 Census, the LFS must connect the 1986 blocks with the 1981 blocks. Permanent numbering would facilitate the matching and reduce the costs of such an operation.

**The preparation of collection maps** by the Geography Division could be a less costly operation if such a program were developed. Could the costs of collection be reduced? Could the regional offices define the enumeration areas themselves through the grouping of blocks? All of these questions must be examined before such a program can be proposed.



# COMPUTER SYSTEMS TO SUPPORT CENSUS GEOGRAPHY

GORDON DEECKER, RON CUNNINGHAM AND KAROLE KIDD

INFORMATICS SERVICES AND DEVELOPMENT DIVISION  
STATISTICS CANADA

## 1. Introduction

The backbone of any modern census operation is the geography system developed to support it. This fact is noted by Gordon (1975) who says that "a census must have a spatial framework for which data are to be gathered, tabulated and reported". Marx (1985) summarizes it as follows: "The success of a census rests not only on how well we collect the data, but also on how well we link those data to geographic areas". Tomasi (1985) identifies three major geographic tools involved in the census support function. They are:

- Maps
- Address reference files
- Geographic reference files

**Maps** relating to census geography include: reference maps which show the geostatistical framework; collection maps which are guides to the distribution and collection of the census questionnaires; and thematic maps which display specific themes or data variables for a given set of geostatistical units.

**Address reference files** contain digital representations of street networks and associated features, along with address ranges for each side of a street between consecutive features. These files are called GBF/DIME files in the U.S. for the 1980 Census, and are referred to as Area Master Files (AMF) at Statistics Canada.

**Geographic reference files** contain, in digital form, the relationships between the various components of the hierarchy of geostatistical areas. The Master Reference File (MRF) in the U.S. and the Census Geographic Master File (CGMF) in Canada are examples.

The geography and cartography components of the census are supported by computer systems which are necessary for the processing of the large and complex sets of spatial data involved in any modern census.

In this paper we shall look at the spatial systems that were put into place for the 1971 Census and the pressures that have come to bear on these systems. The changing nature of the statistics-

gathering process and of statistics-users' demands indicate that significant change in these spatial systems is inevitable. We will attempt to show how these demands can be converted into opportunities for improvement.

## 2. Census Geography Systems at Statistics Canada

The development of geography systems for support of the Canadian census began in the late 1960s when spatial information system technology was still in its infancy. Ion (1969), Podehl (1971) and the GRDSR manual published by Statistics Canada (1972) document the systems that are essentially still in use today. During the early 1970s the CGMF data files were developed to complement the urban data coverage, and to provide uniform coverage for the entire country for selected geostatistical areas.

The Geographically Referenced Data Storage and Retrieval (GRDSR) system was developed for the 1971 Census as a means of retrieving census data for any user-specified area, whether or not it respects standard geostatistical areas. The retrieval process is based on a series of x-y centroids, defined at the block-face level in AMF areas, and at the EA level in the rest of Canada. These centroids are linked to the appropriate information in the census data base. When a user requires information for a specific set of areas, he/she outlines them on a map. These query areas are converted to digital form and GRDSR identifies all centroids. The statistics are then tabulated in a report.

An AMF references every street, address range, block-face, and centroid coordinate in cities with a population of 50,000 or more. Also itemized are other features such as railroad tracks, rivers and municipal boundaries. The AMFs were designed to enable data retrieval for non-standard areas and are now also being used to produce collection maps for some major urban centres for 1986. The background, content and applications of the AMF are described by Parker, Yan, and Kidd (1985).

The main purposes of CGMF are: (1) to provide a central computer data base of the geographic information for the census, and (2) to quality

assure the geographic data by ensuring the correct aggregation and presentation of census data. Information is coded in three data files: the geographic aggregation data file (EAMF) which contains one x-y coordinate reference for each enumeration area (EA) and a series of codes to identify all the geostatistical areas in which the EA falls; the attribute data file (ANAM) which describes the Standard Geographical Classification and official name for each area, and supplementary information such as land area and population count from the previous census; and the boundary data file (ABND) which defines the boundary polygons for certain geostatistical areas.

Since 1976, census geography systems have been operating in a mode of "minimal change", and research on new systems has been limited in scope. Refinements and enhancements are minor relative to the initial research and development efforts, and have come about in response to specific user demands and external pressures. The demands have centred on diversified presentation formats for the geographic data with separate computer programs developed as the requests were made.

In summary, from a single GRDSR system, we have grown to an eclectic set of systems that have been developed to overcome specific problems or to respond to specific enhancement needs. Collectively, these provide a "spatial information system" that is tied together by the AMF and CGMF data files.

### 3. Pressure Points

When systems are left in place for 15 years, there are many pressures brought to bear on them. These are caused in part by technological developments and in part by changes in the application itself. In order to discuss the pressures that are associated with the census geography systems, we have defined the following categories:

- New products
- New clients
- New systems
- New ideas
- New data
- New joint ventures

These types of pressure points will be examined one at a time, with a current example for each being described in detail.

#### 3.1 New Products

Demand for new types of geographic products has taxed the limits of some of the existing

systems, necessitating expansion of the data bases and increasing the need for changes in data structures. Some of the recent products requested include:

- Forward Sortation Area (FSA)/postal code maps
- emergency response maps
- tourism data modelling
- block-face data maps
- multiple symbol maps

The Trillium Data Group was given a contract by the Ontario Ministry of Health to produce an ambulance dispatch system for the Halton-Peel Central Ambulance Dispatch Services. Trillium subcontracted to Statistics Canada the work of extending the AMF in Halton and Peel counties and producing reference maps and a street index for that area. In order to satisfy their needs, however, an increased number of road types had to be defined within the AMF structure.

#### 3.2 New Clients

As part of the 1986 Census cost-recovery effort, emphasis has been placed on creating by-products that can be produced in a cost-effective manner and sold to external clients. Within the past 18 months, a number of clients in the private sector have approached Statistics Canada to obtain hardcopy products and data files that are by-products of the census geography data base. The companies include:

- Gandalf Data Ltd.
- Trillium Data Group
- John Deere and Company
- Canadian Tire Corporation

Statistics Canada has recently provided Gandalf Data Ltd. with a street index, generated from the AMF, for a taxi dispatch application in Metro Toronto. Using these data, the street address of a customer will be input into the system, which will then return the names of the intersecting streets on either side. This will allow identification of the customer's location to the nearest block-face.

In 1984-85, digital files of municipality boundaries for Canada (used for quality assurance purposes in the CGMF - see Section 2) were merged and converted into a form useful for thematic mapping. The file currently contains only the official limits, which do not respect actual shoreline (e.g., the limits go into the water, and often several islands are included in a single polygon). The addition of

shoreline to this file would increase the diversity of its applications, but would also substantially increase the volume of data in the CGMF data files.

### 3.3 New Systems

New systems provide new opportunities for automation of processes and for changes in existing data structures, thus overcoming the problems of excessive data and rigidity. Since GRDSR was developed for the 1971 Census, there have been major improvements in spatial information systems. Typical of these systems are:

- CARIS
- ARC/INFO
- GBF/DIME
- TIGER
- ARIES III

The ARC/INFO system, developed by Environmental Systems Research Institute, is a state-of-the-art geographic information and data base management system. It permits various forms of input, manipulation, user query, and output of geographic data which are not possible with current Statistics Canada systems. It allows for the subdivision and manipulation of geographic data by rectangular "tiles" and would permit the addition of a topological structure to the AMF.

Clayton (1980) describes how image analysis systems are being used, in working with satellite data, to create land-use classification schemes, to detect land-use change on the edge of metropolitan communities, and to derive population estimates in areas undergoing substantial growth. The ARIES III system is now being used by Statistics Canada in the analysis of agricultural data.

New systems are being touted as tools to increase productivity and decrease costs. If these systems are substituted piecemeal into the overall system process, the net result will be an increase in the number of diverse computer systems that need to be maintained.

### 3.4 New Ideas

Throughout the life span of a system, new ideas for products and applications are generated. These ideas in turn can create pressures to modify the basic data structures of the system in order to improve its power and flexibility. Notions currently being considered at Statistics Canada include:

- postal code integration with the data base
- addition of block topology to the AMF
- integration of AMF and CGMF data files
- EA mapping
- input, storage and output of cartographic data by standard grid units
- linkage with the Electronic Atlas project

Currently, streets and other line features in the AMF are represented in a form that contains no information about the adjacent areas (e.g., blocks, geostatistical areas). Other geographic files represent geostatistical units in a polygon format which does not permit identification of the surrounding areas. The addition of a topological element, as in the TIGER system, would permit a boundary or street segment to be described as a "1-cell", with associated information about its junctions with other nodes (0-cells), and the areas (2-cells) that lie on either side. This would make for straightforward generation and aggregation of areas, as well as other geographic, cartographic and analytical operations.

While the addition of topology would add significant value to the AMF data files, it would mean a modification of one of the basic building blocks of the spatial information system that supports census geography.

An "Electronic Atlas" is being developed by the Surveys and Mapping Branch of Energy, Mines and Resources Canada (EMR), to provide increased flexibility in the manipulation, analysis, and creative use of National Atlas information. The Electronic Atlas does not simply store existing maps, but also stores both positional and attribute data in such a way that creation of new maps is possible through the interactive manipulation, analysis and display of the data. EMR is interested in interfacing census data and geographic boundaries with the atlas data files. The effort to develop new presentation formats for census files designed mainly for quality assurance purposes may lead to programming and other methodological problems.

### 3.5 New Data

Most joint venture agreements to date have involved extension of the AMFs. One such arrangement has recently been made with the city of Woodstock and the Government of Ontario to create an AMF for Woodstock. Bradley (1985), in an internal memo, points out that this success could be repeated in

"hundreds" of other communities. This would add a substantial amount of new data to an already overloaded system.

Additions to the volume of geographic data cause pressure on processing requirements, at times exceeding the limits or reducing the efficiency of some processing and plotting programs which were designed for significantly smaller data bases. This type of situation has usually been circumvented by the "quick fix" solution, which leaves much to be desired in terms of efficiency.

The following list summarizes some of the pressures that generate new data:

- The quantity of AMF data has more than tripled since 1971.
- With agreements with Woodstock and other agencies it could double again.
- Municipalities of 20,000 - 50,000 population want their own AMFs
- Increased accuracy implies more data.

Table 1 shows the growth in the number of records in the AMF data base for each of the censuses since 1971. It is anticipated that by 1991 the number of records in the data base will increase by 500% compared to 1971.

The National Research Council has demonstrated that the AMF is an effective data base for police departments. The Geographic Resource Allocation Software System (GRASS) described by Arnold (1985) has been well received. Currently, a task force is recommending its use in 120 communities across Ontario. Many of these communities have less than 50,000 population. Thus, there will be pressure on Statistics Canada to extend the AMF base to include these areas.

In 1981, the city of St. Catharines contracted with Statistics Canada to upgrade the quality of the AMF so that it could be used by the Engineering Department. In the main, the upgrades were done to obtain increased node locational accuracy, which enables calculations involving street length to be used in city planning studies. The net result was to increase the size of the AMF for St. Catharines by 300%.

### 3.6 New Joint Ventures

Joint ventures with other agencies lead to changes in structure, system, and processing requirements. Typical of the joint ventures

under consideration or already agreed to are ventures with:

- EMR
- Cambridge
- Woodstock
- IST (CRAR)
- Metropolitan Toronto
- Burnaby
- Winnipeg
- Calgary

As part of the Electronic Atlas project sponsored by EMR, Statistics Canada, through the GCG subdivision, is providing programmer support for further development work, so that common objectives and simplified interfaces for data exchange can be achieved.

Since 1979, the city of Winnipeg has used Statistics Canada's AMF as a source of street address information for dispatch of all emergency vehicles (fire, police, and ambulance). As well, the AMF is used to assign coordinates to each property parcel, permitting spatial analysis for planning purposes. Winnipeg now performs all updates to their files and Statistics Canada uses these digital updates. All plotting is currently performed by Statistics Canada.

Exchanges such as this one with Winnipeg provide Statistics Canada with improved quality data. They also place limits on significant system changes that can be accommodated without violating the joint agreements. If data and interchange formats are modified, then systems at Statistics Canada and Winnipeg must be amended.

### 3.7 Other Issues Impacting On Systems

There are also issues of a political or economic nature that affect the future of geography computer support systems. One such example is a decentralization of the work-load. Regional offices would like more input in the definition of geostatistical areas, including enumeration areas and census tracts. This trend to regionalization of work-load may require changes to the systems.

The cost of software is increasing as a percentage of total system costs. This leads to an increased emphasis on sharing of software and on development of more general purpose systems.

**Table 1. Area Master Files**

Year	Number of records	% increase
1971	293,000	
1976	590,000	101
1981	721,000	22
1986	1,000,000	39
1991	2,000,000	100

Recent changes in technology will play a major role in shaping the future of geographic systems. The decreasing cost of hardware allows more automation of certain processes by the use of sophisticated equipment, e.g., scanners. The increased availability of micro-computers will mean more decentralization, i.e. more analysis of statistics by individual users, with a resultant demand for specialized micro-based geographic software, and for positional and statistical data in diskette form.

### 3.8 Summary of Pressures

Many of the pressure points we have discussed are intertwined. New clients want new products which require additional data to be added. New ideas and new products result in changes to data structures. Joint ventures result in new ideas and new data.

Clearly, the roles of the files maintained for the census geography are changing. AMF data files are being used throughout the country as a base for municipal spatial information systems. CGMFs, previously used for quality assurance, are now being looked to as a base for thematic mapping of census data. However, the systems that support these files have failed to keep pace with the changing demands.

The increased speed and power of the current computer hardware has made it practical to consider applications, which, when GRDSR was first designed, were not economically feasible. One such application is mapping at the EA level. If this notion is accepted, it will result in the integration of data currently maintained separately in the AMF and CGMF data bases, and bring with it a new set of demands for data structure changes.

### 4. Related Studies

An analysis of the literature since 1983 indicates a common ground between STC problems and those faced by other systems.

Peuquet (1984) notes that "the rapidly expanding range of data ... and need for their combined use ... have revealed two severe problems ... narrowness in the range of applications ... and unacceptable storage and speed efficiency for current and anticipated data volumes".

Data problems play a significant role. White (1983) says that "sharing data among computer-assisted cartographers is so problematic it is generally avoided". This notion is supported by Tomlinson (1984) in his keynote address to the International Symposium on Spatial Data Handling in Zurich, Switzerland, when he says that, "Of immediate concern is the ability to exchange data between bases; ... for example, to transfer data between two of the major geographical bases in Canada (CANSIM and CGIS), it is still more cost-effective to plot and redigitize the data rather than to effect digital transfer".

Problems with geographic data handling stem from three main causes, which Pequet (1983) summarizes as follows:

- "boundaries tend to be very convoluted and irregular"
- "data in digital form tend to be incomplete, imprecise and error-prone"
- "spatial relationships tend to be fuzzy or application-specific, and the number of possible spatial interrelationships is very large"

It is apparent that the size of the data base is also a fundamental problem. Tomlinson (1984) suggests that "few of us have any idea how large data bases will become or need to become". Currently, USGS is in the process of digitizing approximately 1,800 sheets for the TIGER system. In Canada, about 1,000 sheets at 1:250,000 scale are being encoded. However, plans for the future are much more ambitious. USGS plans to digitize 55,000 sheets, a process that would derive  $1.5 \times 10^{15}$  pixels of information.

Since the problems are common, we should also look at the recommendations suggested for future developments to see if we can adapt a common ground. Collins et al. (1983) suggest there are three areas of improvements:

- conversion to relational data base systems;
- enhancement of information content by the addition of topology data;
- improvement in human interfaces with systems.

Nuttall and Korenstein (1985) outline the blueprint for a Geographical Referencing and Information System (GRIS) for Metropolitan Toronto initiated in 1984. This provides valuable insights because of the use made of both CANSIM and AMF data by the Metropolitan Toronto planning agency. The intention is to link five data banks, containing a total of 964 billion bytes of data, via a common spatial framework.

The GRIS is designed to serve six basic functions including:

- "permit data drawn from sources with different geographical identifiers to be filed together in a single source with a common geographical identifier";
- "assist data analysis under geographical corridor or catchment area constraints";
- "produce cartographic maps at any scale".

These functions are also basic requirements for Statistics Canada spatial systems. Given that the AMF data are a vital building block in the Toronto system, it is anticipated that joint discussions on problems and their resolution would provide valuable guidelines for future Statistics Canada spatial systems.

## 5. Responses to Pressure Points

There are three possible responses to the pressure on the existing systems that support census geography:

- freeze existing systems;
- enhance existing systems;
- change to new systems.

### 5.1 Freeze Existing Systems

Freezing existing systems is identified as the response which authorizes only the minimum amount of work required to keep current systems operational. In the view of the authors, this response is unacceptable for many reasons, including:

- system problems in current systems;
- it would cancel many other geography initiatives;
- market pressures;
- excessive processing costs will result if data volumes expand;
- credibility.

Boisvenue et al. (1983) in a review of spatial systems noted two major problems. First, many of the programs that have been created over the past 12 years have been written in different programming languages with hard coding of the options with respect to data storage and processing alternatives. The net result is that these parts of the system will have to be rewritten, if only so that the system is maintainable in production mode. Second, the report notes that our ability for producing data from current and past censuses on a stable geographic frame is not developing at the speed expected by clients.

### 5.2 Enhance Existing Systems

Enhancement of existing systems is identified as the response which authorizes increased automation of individual processes by the replacement of manual procedures by automation. This is accomplished either by the extension of existing systems or the development (or acquisition) of an additional system.

Key factors in the discussion of enhancement as a valid response to the pressure points are the notions that:

- some systems cannot be enhanced but must be replaced;
- enhancing individual systems does not readily permit an integrated approach;
- additional spatial relationships cannot be introduced into current systems by way of enhancements.

Boisvenue et al. (1983) comment that "most data files were developed for a particular application with little thought to their more universal applicability". While this is a somewhat harsh criticism for a system that has endured for 15 years, it points out the difficulty of enhancing systems to extend the application range.

### 5.3 Change to New Systems

Change to new systems is identified as the response which authorizes modifications or replacements to systems that exceed current minimum requirements based on today's estimates for needs in 1991.

Change is always a calculated risk, and usually involves a dual effort. Current systems must be maintained as a fail-safe mechanism while new systems are developed and proven operational.

With respect to census geography computer systems, there are ways to minimize the risks being taken. Chief among these is what might be termed the "leap-frog" approach. Table 2 identifies the various system advances that have been made by U.S.A. and Canada for the past 20 years..

Table 2. The "Leap-frog Approach"		
Census	U.S.A.	Canada
1960	FOSDIC Address coding guide DIME	Street indexes
1966		
1970		
1971		AMF Non-standard area retrieval
1980	GBF/DIME	Computer-assisted collection mapping
1981		
1990	TIGER	
1991		

As can be seen from the table, each country has gained valuable insights from the work done by its neighbour. Ideas and procedures have been adapted to the differing needs of each census bureau. However, the substantial research that produced the ideas and procedures has been shared rather than duplicated. The U.S. Bureau of the Census, by its willingness to send delegates to discuss and demonstrate the TIGER system, has shown the desire to share the research that has preceded the development of the system. This knowledge, reshaped to the Canadian context, will significantly reduce the risk associated with authorizing major changes in the computer systems to support the 1991 Census.

## 6. Summary

Geography, cartography, and their computer support systems are basic building blocks of the census program. The current systems have been

in place since 1971 and are bursting at the seams. White (1983) talks about the aphorism: "It ain't what you don't know that hurts, it's what you know that isn't so". With respect to the census operation this is not quite true: census credibility is put at risk in either case. We should not wait until political pressures cause change, as was the case in the U.S.A. The 1980 Census, according to Brugioni (1981), was termed "inaccurate" and "an exercise in futility" by P.M. Klutznick, then Secretary of Commerce.

The objectives of today have not changed much since they were defined when GRDSR was developed. Fellegi (1967) noted that: "It is important that developmental work get underway towards general tools to achieve economies to enable us to deal with massive volumes of data and build in important elements of standardization". In the same document he urged fiscal restraint: "We shall have to make sure that future growth will be controlled and well coordinated and that it will be achieved by efficient utilization of the financial and manpower resources".

What has changed since GRDSR was developed is the size and scope of the problem. Data volumes will have increased 500% by 1991. Topological relationships are being considered for addition to the reference files. In addition, a multiplicity of spatial relationships are being evaluated for inclusion in the data bases in order to extend the potential range of applications.

Brackstone (1983) makes two very important points when he says that "the scale of the census provides opportunities for recovering investments in automation ... its crucial importance implies that technology has to be tried and proven before incorporation". The implication is that "bridge financing" required to support initial research can be recouped during production. But, at the same time, systems must be ready earlier than 1991 in order to prove their reliability before incorporation in the census production process.

The proposal is, therefore:

- Use the results of the TIGER research to minimize duplication of effort.
- Use the available lead time as an opportunity for a thorough system review.
- Let's not try to do it all at once, but
- LET'S START NOW.

## ACKNOWLEDGEMENTS

The authors would like to acknowledge the ideas and comments from S. Wituk and other members of the Geocartographics Subdivision and from D.R. Bradley and his staff of the Geography

Division. Thanks are due also to the members of the GCG Centre Operations Unit who were able to read our handwriting and input the document to the word-processing systems at Statistics Canada.

## REFERENCES

- Arnold, John, 1985. "A Police Application of AMF - The Geographic Resource Allocation Software System (GRASS)", paper presented at the 1985 Conference of the Urban and Regional Information Systems Association, Ottawa.
- Boisvenue, A., R. Bradley, M. Decarufel, R. Graves, A. Porter, M. Renaud, and J. Yan, 1983. "Geography Division Spatial Systems Review" internal report, Statistics Canada.
- Brackstone, Gordon J., October 1983. "The Impact of Technological Change on Census-taking", *Proceedings of the VIIIth Inter-American Statistical Conference*, Buenos Aires.
- Bradley, D.R., October 3, 1985. Internal memo.
- Brugioni, Dino A., 1983. "The Census: It Can Be Done More Accurately with Space-age Technology", *Photogrammetric Engineering and Remote Sensing*, Vol. 49, No. 9, pp. 1337-1339.
- Clayton, C., and J.E. Estes, 1980. "Image Analysis as a Check on Census Enumeration Accuracy", *Photogrammetric Engineering and Remote Sensing*, Vol. 46, No. 6, pp. 757-764.
- Collins, Stanley H., George C. Moon, and Timothy H. Lehan, 1983. "Advances in Geographic Information Systems", *Proceedings of the Sixth International Symposium on Automated Cartography*, Vol. 1, Ottawa-Hull.
- Fellegi, I.P., and J.I. Weldon, 1967. "Computer Methods of Geographical Coding and Retrieval of Data in the Dominion Bureau of Statistics, Canada", *Proceedings of the Social Statistics Section, American Statistical Association*, pp. 53-57.
- Geography Staff, 1982. "Geography and the 1981 Census of Canada (Geography Series)", Working Paper No. 2 - GEO 82, Statistics Canada.
- Gordon, Marvin, 1975. "Cartography for Census Purposes", *World Cartography*, Vol. 13, pp. 3-18.
- Ion, R.J., 1969. "The Geographic Basis of the DBS Geocoding System for Urban Areas: An Overview", Analytical and Technical Memorandum No. 3, Dominion Bureau of Statistics, Ottawa.
- Marx, R.W., 1985. "Developing an Integrated Cartographic/Geographic Data Base for the United States Bureau of the Census", paper prepared for presentation at Statistics Canada.
- Nuttall, Gerry and Gary Korzenstein, "Evolution of a Regional Information System for the Metropolitan Toronto Planning Department", *Proceedings of the 1985 Conference of the Urban and Regional Information Systems Association*, Ottawa.
- Parker, J.P., J. Yan, and K. Kidd, 1986. "Area Master Files: A Better Way to Serve Census Needs, But How Far Should We Extend the Coverage?", *Proceedings of the International 1991 Census Planning Conference*, Statistics Canada.
- Peuquet, Donna J., 1984. "A Conceptual Framework and Comparison of Spatial Data Models", *Cartographica*, Vol. 21, No. 4, Toronto.
- Peuquet, Donna J., 1983. "The Application of Artificial Intelligence Techniques to Very Large Geographic Data Bases", *Proceedings of the Sixth International Symposium on Automated Cartography*, Vol. 1, Ottawa-Hull.
- Podehl, W.M., 1971. "An Introduction to the Generalized Tabulation Programs STATPAK and STATAPE and Their Language TARELA", paper prepared for the Census Data Access Program Workshop, Statistics Canada.
- Statistics Canada, June 1972. "GRDSR: Facts by Small Areas", Introductory Manual, Ottawa.
- Tomasi, S.G., 1986. "Geographic Support for the 1990 Decennial Census", *Proceedings of the International 1991 Census Planning Conference*, Statistics Canada.
- Tomlinson, R., 1984. "New Frontiers in Cartography", Keynote Address to 1984 International Symposium Spatial Data Handling, Zurich, Switzerland.
- White, Marvin, 1983. "Tribulations of Automated Cartography and How Mathematics Helps", *Proceedings of the Sixth International Symposium on Automated Cartography*, Vol. 1, Ottawa-Hull.

# RÉSUMÉ OF DISCUSSION A COMPARISON OF AMERICAN AND CANADIAN PLANS FOR THE CENSUS IN THE 1990s

SID WITIUK

INFORMATICS SERVICES AND DEVELOPMENT DIVISION  
STATISTICS CANADA

## Introduction

The first presentation by Mr. Silla Tomasi gave us a very brief, but effective, overview of what is being done in the U.S. for the 1990 Census with respect to the TIGER geographic support system. The three talks given by Statistics Canada staff will be considered as a single presentation that provides a thumb-nail sketch of what is possible in the way of geographic and cartographic support for the Canadian census of 1991. I will draw some comparisons and contrasts between the two censuses, make some summary comments, and then try to indicate the key issues for discussion.

Let us first look at TIGER, in terms of its **goals, business case and prognosis for success.**

## 1. TIGER System

### 1.1 TIGER Goals

The U.S. Census Bureau wants to have: (1) a complete integration of existing files; (2) a complete theory-based and vertically-integrated **data base** (that is to say, all the information, geographic and cartographic, will be collapsed into one level); (3) a nationwide block program to facilitate collection, processing and dissemination, and (4) their geocoding coverage extended to reduce costs through increased use of mail-out and mail-back.

### 1.2 TIGER Business Case

How did the U.S. Census come to be in a situation to propose all these changes? They built an effective business case. Our colleagues in the U.K. wanted to know how to sell such a program. It's straightforward: all you have to do is promise to do it all **digitally**, do it **sooner** than previous censuses, do it **better** and do it **for less**, and you will probably get the necessary approvals and get the budget you need. The question that remains is: "Will they really do it?"

## 1.3 TIGER Prognosis for Success

I recently had the opportunity of participating in the evaluation of the TIGER system as a member of the TIGER Evaluation Team. The team essentially viewed TIGER as a major managerial problem and not as a major technological problem. TIGER involves the coordination of massive volumes of information, almost 400,000 maps (5 or 6 copies each) that must be produced in 8 to 12 months. It is a very large job, but from the technological perspective we felt that it was achievable and that it should succeed.

The evaluation team identified a number of important pre-conditions: (1) if it is carefully managed; (2) if it is not affected by cuts; and (3) if it is tested early. One of the greatest difficulties encountered when undertaking a massive transformation to automation like that being attempted in the U.S. is that the final product is not seen until all the previous stages have been achieved. They are now working on the first stage, and preparing for the second and third stages. It is likely that they will find problems in the later stages that have ramifications on operations in the earlier stages. For that reason, it is imperative to test TIGER early and make sure those problems are found as soon as possible. There is also a need to improve coordination with the USGS. While there is an exceptionally cooperative working relationship with the USGS, there are always things that need attention (at the detail level). They are working at improving coordination, but some additional strengthening was seen as necessary. In addition, the world-wide general shortage of people specialized in geographic information systems is likely to have some effect.

## 2. GCSS91 System

Looking to the Canadian presentations, and giving the system elements the acronym (that I'm sure no one will want to permanently adopt), the

GCSS91 system (the Geocartographic Support System for the 1991 Census), let us again consider: "What are the goals, business case and prognosis for success?"

## 2.1 What Are the Goals of GCSS91?

The goals are: (1) increased integration of existing files; (2) improved structures for **interrelated** geographic and cartographic data; (3) a staged block program, probably tied to an Area Master File coverage; (4) extension of the geocoding coverage in response to both internal and external demands; and (5) a staged mapping capability which will also be tied to the AMF and the availability of cartographic files in the various sources that were discussed by Mr. Yan.

The current thinking is that a staged approach is appropriate. Total coverage of Canada is not, therefore, likely before 2001.

## 2.2 What Is the Business Case?

Obviously, if one takes a staged approach, the business case is not as overwhelmingly favourable as if a single-step approach is attempted. At the same time, the level of risk is significantly reduced. The level of automation will increase in measured steps. There will be increases in timeliness, increases in quality (consistency, in particular), and (through a staged market-driven approach) decreased unit costs as the volumes go up and the level of automation increases. Actual savings will not be as visible because those savings will be applied to the improvement of the system throughout the process.

## 2.3 What Is the Prognosis for Success?

The extension and enhancement of the Area Master File is mainly a resource availability problem and not a major technological problem. The key questions are: (1) "What is the fund of money that is available?"; (2) "Who will participate?"; and (3) "How can you keep those costs down?". The system integration problem is primarily a time-tabling problem. The block program is only at the stage of investigating the costs and benefits and we must await the judgement of that analysis.

At this point in time, each of these initiatives is, in my view, a **reasonable** step to take.

## 3. Comparisons

Let us make some comparisons between the Canadian and American approaches. The **similarities** generally tend to be in terms of objectives. Both agencies are increasing the level of automation and integration and are moving, at differing rates of speed, towards theory-based structures, putting in place a national block program, extending geocoding, supporting individual addresses and promoting of digital mapping. The major **difference** is in terms of approach. Whereas the U.S. Bureau of the Census is seeking **total** automation and **total** integration, Canada is (1) considering a **staged** approach, (2) carefully prototyping subsystems, (3) moving to a production system for a measured amount of territory, and then (4) moving to a broader coverage in discrete steps.

Canadian file structures have tended to evolve through time in relation to pragmatic constraints and decisions, while the DIME/TIGER structure is fundamentally based on graph theory and topology. Similarly, for the national block program, the proposed Canadian approach is staged and market-driven as opposed to the single-step approach of the U.S. Census Bureau.

## 4. What Is the Bottom Line for the 1991 Census?

We have seen that while common objectives are shared by the two censuses, they have different approaches and different timetables. Both agencies continue to share experiences with each other, and in this case, given the magnitude of change, proposed by the U.S. Census, it might be advantageous for Statistics Canada to be **second**. The Canadian Census should be cheaper as the level of **risk** relating to the amount of change will **decrease**, not only because of the development of theory, concepts and systems, but also because of the availability of cheaper computer resources and the like, and the general availability of cartographic data in digital form through time. The key point is that, unlike in the U.S., in Canada there is a residual need for some **bridge financing** to make the transition to automation. The earlier the start, the earlier the benefits of these files can be realized, and the better the ultimate benefit/cost ratio will be. One of the major achievements by the U.S. Census was to change their monetary resource consumption curve from a classical census pattern to one which involved **front-end loading** of some of the costs, so that they could build the cartographic data base much earlier. In Canada, there is a similar need.

to allocate the money needed for research and development early on, so that a minimal coverage can be established and then the actual market forces will drive the data base and the coverage further out. The research and cost-benefit studies should, as a consequence, begin as soon as possible.

## 5. Key Issues

What are the main issues to be considered, both now and in the coming months? The first is to **set realistic goals** for 1991, and second, to determine **how much risk** we wish to assume or must accept. Third, we need to determine the **time frames** to attain various objectives, and fourth, determine how to go about obtaining the **funding** we need to maintain and increase coverage for digital cartographic files. Fifth, we want to maintain and increase the amount of internal, national and international **cooperation** that can facilitate many of these initiatives. Sixth, **decentralization** is a generally accepted goal that is subject to a number of constraints which prevent it from becoming an instant reality. Decentralization needs to be examined both in terms of its possible role in gathering documentation (much as they are doing at the U.S. Census) and also in relation to the establishment of collection assignments. A seventh key issue is how to derive an **efficiency measure** relating to increased automation. If one draws a graphical comparison (see Figure 1) of progress over the years, it is clear that the Canadian Census has been much more conservative than our colleagues and counterparts from south of the border. We have tended to move in steps that have been less ambitious. I feel that it is time for another major change in terms of systems technology, and in terms of geographic infrastructure. This would be followed by a more **measured** step in 1996. In this manner, and without trying to undertake the **monumental** change that is now being attempted by the U.S. Bureau of the Census, we should be able to reach a similar level and extent of automation by the decennial census in 2001.

### 5.1 Historical Perspective

The technology for spatial information systems has been evolving since the late 1960s

and early 1970s (see Figure 2). As with many technologies of the period, it went through the **start-up** process, the "**gee whiz**, we can do everything", through, "we have really blown it". (This is across all geomatics applications, not particularly the case for census applications.) We "tried again" and had a few minor problems. A threshold has been achieved that makes it easier to proceed than in the past.

### 5.2 How Have Things Evolved?

We can see from Table 1 that the **mail-out/mail-back** methodology in the United States in the late 1960s generated the need for the **address coding guide**. The **automated geocoding** implied by this approach required **topology** and led to the DIME system. The **specialized retrieval** implied **coordinates** which led to GRDSR and eventually resulted in DIME becoming GBF/DIME. It appears that computer-assisted **collection mapping** is now leading to a **national digital cartographic data base**.

Table 1. Cooperative Evolution

Mail-out/Mail-back	⇒	Address Coding Guide
Automated Geocoding	⇒	Topology ⇒ DIME
Specialized Spatial Retrieval	⇒	GRDSR ⇒ GBF/DIME
Collection Mapping	⇒	CACM
National Geographic Data Base	⇒	TIGER

I would like to end this talk by observing that the **risks**, in terms of the **opportunities** we have for **change**, are relatively **small**. Remembering the recent observation by the Director of the U.S. Bureau of the Census, Mr. Keane, "that the greatest risk of all is not to accept any risks", it seems appropriate for us to take those small risks and thereby profit from the many opportunities for increased automation.

Figure 1. Rate of Change in Technology

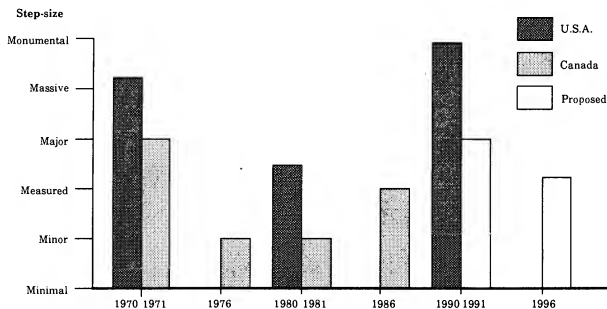
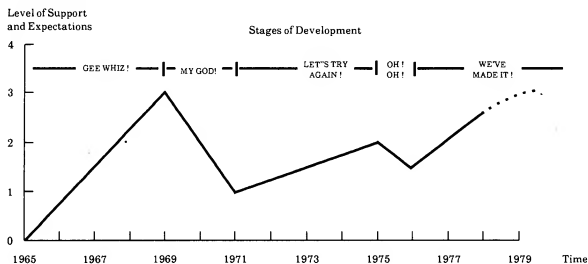


Figure 2. The Development of Geographic Information Systems in Canada



Source: Taylor, D. R. F., Editor, Recent Trends in Geographic Information Processing in the National Capital Region, Department of Geography, Carleton University, Ottawa, May 1978, p. 72.

## **Résumé of the Question Period**

**Mr. Tomasi's** presentation was followed by a question period where several issues related to censuses in the 1990s were discussed. Two in particular concerned the Topologically Integrated Geographic Encoding and Referencing (TIGER) system.

### **Interdepartmental Cooperation**

The first subject was a general talk about the potential benefits of cooperative arrangements between mapping agencies and the Census Bureau.

The U.S. Census Bureau (USBC) is developing the TIGER system for 1990. However, the development of such a system in a short lapse of time requires resources and equipment not within the USBC's immediate reach.

An agreement with the U.S. Geological Survey (USGS) was struck. An initial cartographic data base will be produced by USGS. The USBC during the 1990 Census will collect information which will be used to update the cartographic data base.

Four mapping centres of the USGS centres will work exclusively on the U.S. Census program. This cooperative arrangement is beneficial in terms of equipment utilization and effective use of specialized resources. A further advantage is the improvement in consistency in geographic product emanating from different organisations.

In Canada, Statistics Canada and Energy, Mines and Resources will work together in order to create a national digital data base by 1990. Cooperative efforts also exist between different levels of government (federal, provincial, municipal) to produce more current maps. This updating is generally done jointly by the concerned parties.

Another possible benefit of this cooperative effort is the production of Area Master Files (AMFs) for selected urban areas of less than 50,000 population. This extension of the AMF inventory would represent a significant improvement and cost very little to each party concerned.

### **Role of the Regional Offices in the TIGER System**

The second subject discussed was related to the role played by Regional Offices (ROs) in developing the TIGER system.

ROs had major input in the analysis leading to the development of the TIGER system. They were consulted to determine the major faults occurring in the 1980 Census and provided positive input on specific improvements to the system.

The TIGER system is decentralized in the sense that the 12 regions of the USBC are responsible for the collection of source material, the updating of source documents, some digitizing, and liaison with the local jurisdiction and statistical area program. A geographic staff is hired in each RO to undertake these duties.



**SESSION: THE CONTENT OF THE 1991  
CENSUS**

Chairperson: Ian Macredie  
Labour and Household  
Survey Analysis Division  
Statistics Canada

Wednesday, October 9, 1985



## PANEL DISCUSSION THE 1991 CONTENT DEVELOPMENT PROCESS

HENRY KRIEDEL

AUSTRALIAN BUREAU OF STATISTICS

SUSAN MISKURA

U.S. BUREAU OF THE CENSUS

DAVID PEARCE

OFFICE OF POPULATION CENSUSES AND SURVEYS  
UNITED KINGDOM

### Résumé of Panel Discussion

*The first part of this session consisted of a panel discussion on the actual selection process for the 1991 Census content as experienced by census agencies in Australia, the United States and the United Kingdom.*

#### 1. Australian Experience

**Mr. Kriegel, representative of the Australian Bureau of Statistics, commented on the legal authority of the Australian Bureau of Statistics and outlined the consultation process experimented in Australia.**

*In 1975, the Australian Bureau of Statistics Act was adopted giving the Australian Bureau of Statistics statutory authority, and more autonomy. The Bureau has the ultimate authority in determining census content. The Act also created an Australian Statistics Advisory Council. This council, made up of representatives from the public and private sectors, works in association with the Bureau in the content selection process.*

*Since the 1981 Census, the content selection process has been oriented in a way to give all users of census data an opportunity to submit recommendations.*

*This content selection process was repeated in 1986 because the Bureau recognized that the process in 1981 had been successful, and that it created general public support for the census. Mr. Kriegel briefly highlighted key places in the 1986 content collection process.*

*One of the first steps in the selection process involved the placing of advertisements in newspapers across the country inviting the public to recommend topics for inclusion.*

*Next, a series of user consultations took place. Users were asked for their recommendations on questions on the census questionnaire. These recommendations could pertain to changes, additions or deletions of the question(s). Users were also asked to examine the suitability of topics that could be integrated into future census questionnaires.*

*From these recommendations, certain topics were then selected, and a series of field tests were conducted to assess the impact of these recommendations on respondent burden and to verify if the census was the appropriate vehicle for the topic.*

*After these assessments were completed, a paper containing results on preliminary findings was released. This paper included recommendations from the Bureau on topics to be included or excluded on the census questionnaire. It also identified alternative sources available from which data on certain topics could be obtained.*

*Following the user consultation phase, recommendations were then passed to ISAC and then to the government for final approval.*

*The selection procedures should be revised. The result of the 1986 consultation process revealed two areas requiring further improvements: a reduction in cost and the length of time it takes to carry out the process. One possibility may be to ask users for the recommended topics via different monthly survey reports. These survey reports could present a list of different topics on the back cover inviting users to make their recommendations.*

*Whichever method is adopted for the 1991 Census content selection process, the last two censuses revealed that involvement of major "public" users is essential.*

## 2. United States Experience

**Ms Miskura, representative of the U.S. Bureau of the Census,** discussed the consultation process in the United States.

*The content development process and schedule in the U.S. are mainly determined by the Bureau's obligation to the government that topics of the 1990 Census must be identified on April 1, 1987, and the actual questions one year later.*

*A review of the 1980 Census was the first step in content determination for the 1990 Census. This review involved an assessment of data quality and use of the data by a variety of users. These users were consulted through a variety of forums specifically held to discuss these needs. Local public meetings, which are sponsored by the Census Bureau in conjunction with state and local organizations, are the major forums for consultation. They have provided a wide variety of users, from both public and private sectors, with the opportunity to express their critical judgement on the accuracy of the data and to suggest new or modified data elements for the upcoming census. At least one such meeting has been held in every state. Other forums, conferences and the Bureau's own efforts are also major sources of recommendations through which the content of the census is determined.*

*For determining federal data needs, the Census Bureau has organized a council from other agencies through ten Interagency Working Groups (IWGs), and also through the Office of Management and Budget's Federal Agency Council on the Decennial Census. The IWGs are composed of appropriate subject-matter experts from approximately 80 federal agencies. The major objectives of these groups are: to obtain information about the legislative and programmatic use of census data; to identify the 1980 data elements and tabulation that will not be needed again in 1990; to discuss their interest and concerns, in order to suggest new questionnaire wording formats and answer categories for the 1990 Census, and to provide information about the geographic levels for which data are needed. These IWGs categorized their recommendations into high, medium and low priority data needs and whether the data could be obtained from a sample of the population.*

*A general principle governs the selection of subject content for the census. The census must be aimed solely at data that are required*

*to meet well-demonstrated public needs, to fulfill legal mandates, or to carry out governmental programs. Other criteria considered include the need for data on small geographic areas or for small and dispersed groups, the possibility of obtaining these data from sources other than the census, the feasibility of an understandable wording of the question, and the cost of the proposed new topic.*

*The Census Bureau specialists identified an initial selection of candidate questions which then go through a testing program. The main testing vehicle is the National Content test. These tests are designed to provide information on the reliability of the data collected and the ability and willingness of respondents to answer the questions.*

*One of the major objectives of the 1990 Census is to balance the growing demand for new data needs against respondent burden, i.e. the time taken to complete the questionnaire.*

*In being sensitive to respondent burden, the Bureau is studying the feasibility of reducing the number of questions. The easiest way is by eliminating certain questions or by moving some topics from the 100% to the sample form. Another possible alternative is to use multiple sample forms which would allow for collection of new data without increasing the average length of time it takes to complete the census questionnaire. Those strategies, and others, would help to meet the objective of balancing between growing content requirements and respondent burden.*

## 3. United Kingdom Experience

**Mr. Pearce, representative of the United Kingdom,** discussed the consultation process in the U.K. in terms of identifying the interested parties, how and when should these parties be consulted, and finally, how much weight should be given to these user consultations in determining the census content.

*There are several interested parties. The central government needs data for allocating funds, local governments need figures that are reliable for small geographic levels, the academic community requires data for research purposes, and finally, businesses need census data for developing marketing strategies. Without a doubt, census data are useful to various groups in both public and private sectors. In terms of meeting these various needs, it is essential that the process of*

consultation be based on a consultation program.

*There are two levels of consultations: one is an informal process conducted at the grass-roots level in the form of ad hoc meetings, and the other is a more structured and formal process involving a series of consultation sessions with other government departments.*

*The informal process takes on a variety of forms such as ad hoc meetings or a series of informal sessions conducted across the country. This informal process can be done with all users, particularly local governments. In 1981, ad hoc meetings were very effective, resulting in the implementation of new questions on topics such as dwellings, address five years ago, and ethnic origin. Also, there is the Monitor Series which is a publication distributed free of charge to approximately 5,000 persons who could be interested in census data.*

*The formal process is more structured. In 1981, the Office of Population Censuses and Surveys (OPCS) organized a series of meetings with other departments on specific topics and established an advisory committee comprised*

*of directors of statistics of the most interested departments, local government authorities, and academic researchers.*

*After the consultation process, OPCS establishes a census committee and produces a discussion paper on census content including recommendations and alternative sources from which the data can be obtained. The paper is distributed to interested parties for comments before it is presented to Parliament.*

*During formal and informal consultation processes, OPCS has the final authority at recommending questions to be included in the census.*

*One final issue should be addressed. The consultation process is very time-consuming in nature. Unfortunately, the process must be done at least four or five years before the census is taken, because of the need to incorporate the questions in the questionnaire form which must be approved by Parliament several years in advance. Therefore, the consultation process itself needs to be shortened in order to identify quickly the key issues concerning the topics to be integrated in a census questionnaire.*



## PANEL DISCUSSION THE 1991 CONTENT PROPOSALS

JOHN KETTLE

FUTURESEARCH PUBLISHING INC.

FRANK CLAYTON

CLAYTON RESEARCH ASSOCIATES

NOAH MELTZ

UNIVERSITY OF TORONTO

### Discussion

The second part of the session consisted of content proposals for the 1991 Census, from three major census data users.

#### 1. Representative of Futuresearch Inc.

**Mr. Kettle, from Futuresearch Publishing Inc.,** made some suggestions on data for 1991. These suggestions were made not only because the data would be of intrinsic interest to the author, and probably others, but also because of the prospect of correlating these data with other information normally gathered during the census, like age, sex, education, income, occupation. The proposals were grouped under three headings.

#### Health

Although good data are available on the aggregate use of medical facilities, there appear to be no data showing individuals' use of hospital and medical facilities. Both current (last 12 months) and cumulative (lifetime) data would be useful. Suggested items could include visits to doctors, stays in hospital, time off for sickness, stated ailments, etc. Even one of these would be a useful advance.

#### Education

Better data on current educational activities and cumulative level of education would be extremely valuable. Education has a high value for forecasting other variables, such as labour force activity, fertility rates of women, etc., and better information should improve the practice. This information could include part-time as well as full-time education, training, and retraining. For those with postsecondary education, major subject(s) would be useful, particularly for occupational forecasting.

### Work

Many suggestions arise on this matter, even if it is difficult to do the separation between the objective of the census and the Labour Force Survey.

- **Place of Work.** The intent here is to get a better understanding of the rapid increase in workers in rural non-farm areas. Cross-tabulation of place of work by place of residence should explain some mysteries.
- **Work at Home.** About 10% of workers, it is estimated, currently work at home (or from home). It has been estimated that with current technology as many as 40% could work at home. It has also been estimated that by 1995 about 20% will work at home. The question does not appear to be asked in any current survey.
- **Managers by Level.** There are about 1 million managers currently in the labour force - twice as many people as are primary workers, twice as many as are construction workers or transportation workers - but we have no data on what level of management these managers have reached. Since many changes are taking place here, and even more will probably occur over the next 10 years, better data are important. Even to distinguish between senior executives, middle managers, and supervisors would be valuable. The correlation with sex would also reveal changes that are going to become important.
- **Proprietors.** The same sort of points could be made about proprietors as about managers: the information on proprietors is too skimpy. The correlation with age and sex should reveal dynamics that are going to be increasingly important.

- **Information Workers.** It may not be possible to ask questions about the nature of the work performed with any hope of getting consistent answers, but the fact is that about half of those employed currently do what has been described as "information" work, although the category is not included in any survey or census.
- **Small Business.** The same difficulty arises, perhaps with more force, in trying to determine how many people work for small businesses (fewer than 20 employees is a useful category). It used to be possible at least to estimate this number until Statistics Canada brought in new employment series in the spring of 1983; now it is not. Most new employment appears to arise in small business.
- **Underground Economy.** It is obviously impossible to ask people if they are carrying on business "off the books" or in the black or informal economy, but it may be worth raising the issue in this forum in case someone can think of a new way to dig out some useful information. It appears to be a major and growing category of work.
- **Leisure.** Leisure activities now occupy a considerably greater fraction of adults' waking hours than formal work. We are prepared to make a large effort to understand what is happening at work; we should do more to study leisure. Time allocation outside working hours could be listed under half a dozen categories, such as: parenting, home care, sport, exercise, leisure at home, television, reading, gardening, crafts, idleness, out-of-home culture, entertainment, education, training.

## 2. Representative of Clayton Research Associates

According to Mr. Clayton of Clayton Research Associates, data on housing are essential. Users of data from the Census of Canada have, since 1961, come to expect the inclusion of a comprehensive set of housing questions in the censuses conducted at the beginning of each decade. These include questions relating to tenure, unit type, costs, age, type of heating equipment and number of rooms. Much less housing data have been available from the mid-decade censuses. In this regard, the 1986 Census with its comprehensive housing questions will be a bonus.

## Why Collect Housing Data in the Census of Canada?

The housing data collected in the census are needed for a wide range of policy and market-related purposes:

- **Social Housing Policy.** Government policy makers concerned with identifying Canadians with housing problems need data on the ratio of shelter costs to incomes for all households by type and their housing characteristics.
- **Policy With Regard to Maintaining or Enhancing the Quality of the Housing Stock.** Since the early 1970s, government attention is increasingly being devoted to maintaining or enhancing the quality of the existing housing stock.
- **Policy With Respect to Energy Concerns.** During the late 1970s, the types of equipment and fuels utilized to heat and cool homes became of increasing interest to government.
- **Policy With Respect to the Intensification of Use of the Existing Housing Stock.** In older parts of Canadian cities, there is the potential to create additional units in the existing stock, as many large houses are occupied by as few as one person.
- In order to adequately conduct housing market analyses, there is a need for accurate data on household characteristics, their incomes and current housing situations - particularly as new housing demand becomes increasingly fragmented.

The common thread for all these purposes is the requirement for a comprehensive data base linking housing variables with the demographic and socio-economic characteristics of the population.

Cannot these data be collected through sample surveys? Why, for instance, isn't Statistics Canada's HIFE (Household, Income, Facilities and Equipment) survey sufficient? While sample surveys provide useful information, there is still a need for housing data from the Census of Canada. There are three reasons for this:

(1) The need for bench-mark data.

As with other data collected in the Census of Canada, such as total population and its characteristics, there is a need for periodic accurate counts of key housing data such as the total number of dwelling units (both occupied and unoccupied) by tenure and type. Housing completion data are not an adequate substitute since changes constantly occur in the existing housing stock. Moreover, because many uses of housing data concentrate on changes in the housing stock and its characteristics, sample surveys are inadequate since changes in the total stock figures from sample surveys encompass both actual change and sampling error. For example, shifts in the tenure mix of the housing stock since 1981 are not known due to sampling error in available estimates (based on sample surveys).

(2) The availability of small area data.

This is probably the most important attribute of the census with respect to housing data. The Census of Canada is the only source of housing data (as well as a comprehensive set of integrated housing, demographic and socio-economic data) for small geographic areas.

Why is this important?

- (a) Housing markets, by their very nature, are localized (e.g., Ottawa or Halifax), and even within these areas, there are many submarkets. This situation arises because housing is a unique good – it is immobile and has a long life.
- (b) Government policy makers need data on small areas in order, for instance, to identify specific areas with large concentrations of housing in need of major repairs.

The ability to choose specific geographic boundaries is a distinct advantage of the Census of Canada since the data collected are geocoded.

- (3) There simply are not adequate substitute sources for housing data from the census.

Useful housing and housing-related data can be obtained from sample surveys such as the Survey of Family Expenditures, the Survey of Consumer Finance and the very valuable HIFE survey. However, these surveys simply cannot provide the historical data, the reliability and the small area data provided by the census. These surveys have many limitations even for estimates for the subarea provinces, let alone any smaller subareas.

### Is the Census Collecting the Right Housing Data?

The 1981 Census obtained a variety of housing information including:

- the occupied stock by tenure and type of unit;
- dwelling characteristics such as age, type of heating equipment and fuel, number of rooms, number of bathrooms and dwelling condition; and
- shelter costs (for both renters and owners) and the value of owner-occupied dwellings.

Generally, this range of information does meet the needs of most users.

There are a few questions on the 1981 Census questionnaire which, upon reflection, appear to be of marginal use. These include the type of fuel used for water heating and the number of complete and half bathrooms. One also wonders why the length of occupancy was designated as a housing question in 1981.

### What Are the Housing Data Needs from the 1991 Census?

The world, insofar as housing is concerned, is changing. During the 1990s, the level of new residential construction activity will experience a sharp decline as a result of demographic forces. Due to this decline and the aging of the existing housing stock, government and private businesses will increasingly turn their attention to maintaining and improving the existing stock.

At present, there is a large gap in residential renovation statistics which will inevitably grow. Considerable thought should be given to the question of how the 1991 Census might be able to fill this gap. Possibilities include

obtaining information on the extent and types of renovation activity undertaken by home-owners during the preceding 12 months; the amount spent by home-owners on these renovations (by activity) during the preceding 12 months; or more details of the housing stock (e.g., number of windows, presence of a garage).

### **Is there Likely to Be a Conflict Between Historical Continuity and New Subject-matter in the 1991 Census?**

The issue of continuity in census data is a real concern. However, as long as the housing questions on the 1991 Census are comparable in scope to those on the 1981 Census, there does not appear to be a serious problem with data consistency over time. Reliability of existing data (e.g., housing type data from the 1981 Census) is probably a larger concern than is continuity.

### **How to Handle Conflicts Between Unlimited User Demand for Housing Data and Limited Capacity for Housing Questions?**

Again, if the 1991 Census is comparable in scope to the 1981 Census, it appears that Statistics Canada will not experience insatiable pressure for extensive additions to the housing content of this census.

More than being the only adequate source for housing data, it is also the most cost-effective method in the collection of these data.

### **3. Representative of the University of Toronto**

**Mr. Meltz, of the University of Toronto,** presented his comments on improvements on data content for the 1991 Census. My remarks will deal with labour market data by focussing on four topics: premises concerning the census as a source of labour market data; likely labour market issues in the 1990s; and 1981 Census and labour market data; and content considerations for the 1991 Census.

#### **Premises Concerning the Census as a Source of Labour Market Data**

In considering proposals for the 1991 Census, it is necessary to set out the objective of decennial censuses, that is, what can and cannot be done. The objective seems to be to provide an in-depth reference point describing

various characteristics of the Canadian population. From the labour market perspective the characteristics which are unique to the census are socio-economic statistics on the income, employment and unemployment experience of the population classified by detailed occupations involving in 1981 approximately 500 kinds of work in almost 300 kinds of business or service industries. The socio-economic characteristics include: age, sex, amount and type of education, ethnic origin, religion, language, country of origin, period of immigration, place of residence, place of work, etc. No other survey provides as comprehensive a set of data on detailed occupations and geographic areas as the census (Meltz 1982).

However, the breadth of coverage of the population and the depth in terms of detail for specific subjects have a price. The price is that the questionnaire has focussed on a much smaller number of aspects in the labour market than its regular counterpart, the monthly Labour Force Survey (LFS). The LFS has much more depth on questions relating to hours of work, job search, and unemployment. On the other hand, the census provides more detail on education, training and income. While the two surveys are related they also have somewhat different purposes. The LFS is designed as a thorough examination of employment and unemployment experience. But it lacks the ability to provide data on detailed occupations or industries or on small areas within the country. The census provides a detailed reference point.

### **Likely Labour Market Issues in the 1990s**

In order to propose changes (or additions) in content, it is necessary to indicate the rationale for these changes. One perspective on this aspect is the likely labour market issues of the 1990s. What subjects will be of concern in the near future for which the census will be called on to provide an input? Among possible subjects are: the demographic structure of the population and the work-force in general and in detailed occupations and industries; pensions and the labour force activities of an increasingly aging population; other non-work income as well as benefits received at the work place; equal employment opportunity; equal pay for work of equal value; the increasing proportion of part-time workers; public sector versus private sector earnings; unemployment; and immigration.

## **The 1981 Census and Labour Market Data**

The census can provide information on labour supply and on the results of the interaction of labour supply and demand, namely employment, returns to labour (wages, salaries, benefits) and unemployment. The census cannot measure unsatisfied demand for labour.

The census is a primary source of data on demographic structure and is used as the central reference point for many surveys. Data on period of immigration and country of origin are also provided in the census along with information on pensions and labour force activity by age groups.

On the labour supply side, the 1981 Census gives some information on education and training. It does not give a comprehensive inventory of the detailed education and training of the population. It provides the general levels and broad types of achievement.

On the results of the operation of the market, the census provides some aspects of employment, income and unemployment. These aspects, in turn, can all be cross-classified by all of the socio-economic characteristics.

In terms of the possible issues to the 1990s there are several gaps. Such subjects as part-time work, equal employment opportunity and equal pay for work of equal value require additional information on hours of work and wage and salary rates. The census, in contrast with the Labour Force Survey, does not distinguish between the number of hours worked in main jobs versus other jobs (census Questions 39 and 44 versus LFS Questions 18, 76 and 77). Wage and salary earnings are attributed in the census to only one job, whereas there may have been several jobs during the year. Even more serious, the occupation which is reported may not be the one in which the bulk of earnings are obtained.

A second aspect relating to these two subjects is the wage rate, that is, the return per hour or per time period. The 1981 Census does not enable one to calculate a wage rate, a common measure of analysis in labour economics. The weeks of work question (Question 45) deals with 1980 while the hours of work question (Question 39 (a)) relates to the first week of June 1981. Since the time periods are

different and since the wages and salaries could be derived from several sources, it is difficult to estimate a wage rate.

Public versus private sector wage rates and earnings can be only partly dealt with from census data. As indicated above, wage rates cannot be calculated and neither can the extent of unionization. Third, the industrial classification has government as a separate category but government-owned organizations cannot be identified.

While the census has a detailed breakdown on sources of income it does not identify benefits such as vacations, holidays, employer contributions to health plans, savings, etc.

## **Content Considerations for the 1991 Census**

Drawing on the preceding discussion, the following are suggested content considerations. These suggestions are made on the assumption that the purpose of the census remains that of providing an in-depth reference point and that the number of questions available is limited. First, the 1981 questions on hours of work could be expanded to indicate the number of hours in the main job and in other jobs (Question 39 (a)). The same would be required for the status of workers (Question 44), weeks worked (Question 45), and the amount of wages and salaries (Question 46).

A related change is required in Questions 41 and 43. I have always assumed that the cross-tabulations of occupations and wage and salary earnings were internally consistent, that is, the earnings were received for work in that occupation. This was the case up to and including 1961. A careful reading of Question 43 (and similarly for industry in Question 41) indicates that the occupation is the occupation of the week preceding June 3, 1981 which could be different from the occupation from which the earnings were derived in 1980 (Question 46 (a)). Presumably, for most persons the two will be the same, and it is the same for those without work in the last week of May 1981. Nevertheless, there should be an addition to Questions 41 and 43 to indicate whether this was the same occupation and industry from which the main earnings were obtained in 1980. If this were not the case, then the main occupation and industry in 1980 should be requested.

A small but useful addition would be to indicate in Question 45 of the census that part time means less than 30 hours. This is indicated in Question 39 (d), but not in Question 45. Questions 39 (d) and 45 might also indicate that full time means 30 or more hours, as is done in the Labour Force Survey.

Let me mention a slight wording difference between the 1981 Census and the Labour Force Survey. The census, in Question 39 (e), asks whether there was any reason why you could not start work last week. The LFS, in Question 64, asks whether there was any reason why ... could not take a job last week. For consistency, I would prefer to see the census use the term "take" versus "start".

Another consideration is the length of employment with the current employer. A question like LFS 73, on when the person started working for this employer, would add an important dimension to the analysis of earnings.

Consideration should be given to obtain data on wage rates. This is a difficult subject which could be dealt with in a variety of ways. The most direct way is to simply include a question on regular wage rates. An alternative would be to accompany the question on weeks worked with a question on the normal hours of work per week during the period. Because of the problem with the meaning of the wage and salary earnings figure, this could still create difficulties in deriving a figure on wage rates.

On another subject, the fact that 40% of non-agricultural employees are members of unions and that collective agreements cover about half of the work-force suggests that a question be added on union or professional association membership.

Considerations might be given to including a question on employer contributions to vacation, holidays, health, retirement, benefit plans, etc. While this is an important subject, with benefit costs running at 30% to 40% of direct wage and salary costs, it may be difficult to include in the census a precise question on this area.

The final consideration is that of preserving series. The suggestions which have been made would not affect time series since they would be either additions or a subdividing (in the case of main job versus other jobs) such that comparability with earlier series can be reconstructed.

The most important aspect in the labour market data is the classification of occupations and to a lesser extent the classification of industries. In the latter case, the 1980 Standard Industrial Classification will be used in 1991, but it can be regrouped for comparisons with the 1970 SIC used in the 1971 and 1981 Censuses. The crux of the problems is the occupational classification. This is the heart of labour market analysis. Up to 1981, there had been sizeable changes in the classification of occupations with each census. The most major change occurred in 1971 for which only 9 out of 489 classes were entirely comparable with the 1961 Census (Meltz and Stager 1979). Fortunately, the 1981 Census was almost completely comparable. What will happen in 1991?

I urge this group to press for comparability or at least adjustability to the 1971 occupation classification. Professor David Foot and I have just completed a study for Employment and Immigration on the economic determinants of changes in occupational composition of employment. We also assessed the accuracy of past occupational projections. A major aspect of the project was trying to identify which changes in occupational composition were real and which were the result of changes in classification. The major projections for Canada in the 1960s and early 1970s (Meltz and Penz 1968, Ahamad 1969) used classifications which were subsequently superseded when the new censuses appeared.

Human resource analysis and related subjects such as part time versus full time, labour force activity by older persons, immigration considerations, equal employment opportunity, equal pay for work of equal value, etc., all rest on the need for consistency in occupational classification. If there is one single message I would like to convey, it is - please make sure the 1991 occupational classification is comparable with the 1981 and 1971 bases.

---

## REFERENCES

Ahamad, Bill, 1969. **A Projection of Manpower Requirements by Occupation in 1975: Canada and its Regions.** Ottawa: Queen's Printer.

Meltz, Noah M. and G. Peter Penz, 1968. **Canada's Manpower Requirements in 1970,** Department of Manpower and Immigration. Ottawa: Queen's Printer.

Meltz, Noah M. and David A. A. Stager, 1979. **The Occupational Structure of Earnings in Canada, 1931-1975,** Anti-Inflation Board of Canada. Hull, Quebec: Ministry of Supply and Services.



## **SESSION: CENSUS OF AGRICULTURE**

Chairperson: Terry Gigantes  
Resources, Technology and Services Statistics  
Statistics Canada

Wednesday, October 9, 1985



# MAIL ENUMERATION IN THE UNITED STATES

## CENSUS OF AGRICULTURE

CYNTHIA Z. F. CLARK

U.S. BUREAU OF THE CENSUS

PRESENTED BY CHARLES P. PAUTLER

AGRICULTURE DIVISION  
U.S. BUREAU OF THE CENSUS

### 1. Background on the Census of Agriculture

The United States Census of Agriculture is taken to provide a detailed statistical picture of a vital sector of the Nation's economy. The census has generally been taken at 5-year intervals and collects data on land in farms, agricultural production and sales, farm operator characteristics, as well as other agricultural data. These data are used by farmers, government agencies, and private organizations for making decisions, benchmarking surveys, and researching agricultural markets. This paper discusses procedures used in the United States Census of Agriculture that Statistics Canada may want to develop for the 1991 Canadian Census.

Twenty-two censuses of agriculture have been conducted in the United States, beginning in 1840 as part of the decennial census of population. From 1840 to 1950, an agriculture census was taken as part of the decennial census. A separate mid-decade census of agriculture was conducted in 1925, 1935, and 1945. From 1954 to 1974, a census was taken for the years ending in 4 and 9. In 1976, Congress authorized the census of agriculture to be taken for 1978 and 1982 and every 5 years thereafter to coincide with the economic censuses of manufacturing, business, governments, transportation, and construction. The change in reference years increased data comparability and achieved efficiencies in processing operations for these censuses.

The census of agriculture is required by law under title 13 of the United States Code that governs the operations of the Census Bureau, an agency of the U.S. Department of Commerce. The confidentiality of the data is protected by prohibiting the use of the data except for statistical purposes, prohibiting the publication of data identifying any particular individuals, and limiting access to census reports to only sworn officials and employees. The data collection unit of a farm operation used in the agriculture census is established by law. Since 1850, when minimum criteria defining a farm for census purposes were

first established, the farm definition has been changed nine times. The current definition, first used for the 1974 final reports, is any place from which \$1,000 or more of agricultural products were sold or potentially could have been sold during the census year. A place not having sufficient sales to qualify as a farm can qualify on potential sales based on the inventory and production of crops and/or livestock.

The census of agriculture is the leading source of statistics about the Nation's agriculture and the only source of consistent, comparable data about agriculture at the county, state, and national levels. Data from the census are valuable not only to farm operators, but also to the entire agribusiness sector of our economy. Census data, as well as current sample estimates derived from or based on census benchmark data, are widely used for planning purposes by manufacturers servicing agricultural operations, and by businesses involved in the transportation, processing, or distribution of agricultural products to the consumer. Census statistics are used by Congress in developing farm programs and for analyzing the results of such programs. Many national and state programs affecting agriculture are designed or allocated on the basis of census data, such as funds for extension services, research, and soil conservation projects. Individual farm operators can compare their own agricultural activities with totals and averages for their county.

The data collected from farm operations include: acreage, land use, and irrigation; crops including field crops, hay, vegetable products; livestock, poultry, and animal specialties; sales data; characteristics of the operator; use of fertilizer, pesticides, and other expenditures for energy; machinery and equipment and market value of land and buildings. In addition to census data on agriculture, the respondent list of census farms has been used as a sampling frame to collect more specialized data on sectors of the agricultural economy. In the past, such data collections have included the 1979 Farm Finance Survey, the 1979 Farm Energy Survey and the 1979 and 1984 Surveys of Farm and Ranch Irrigators.

The 1982 Census of Agriculture enumerated 2.2 million farms in the United States. These farms had 987 million acres of land and 132 billion dollars of agricultural product sales.<sup>1</sup> The Coverage Evaluation Program for the census estimated that about 91 percent of the farms in the conterminous U.S. were enumerated by the 1982 census.<sup>2</sup> Coverage was much better for the group of farms with sales of \$2,500 or more than for the group of farms with sales of less than \$2,500.

The data from the census of agriculture are made available to users in several different forms. For the 1982 Census of Agriculture, preliminary and final results were available for each county, state, and the United States. Final data were available for the outlying areas of Puerto Rico, Guam, and the Virgin Islands. Preliminary and final reports were available in printed form and on computer tape files. Preliminary data tabulations were also available on microcomputer diskettes in detailed county, state, and national tables.

Four other supplementary 1982 publications are available in printed copy. The Ranking Counties and States publication provides the ranking in order of importance for selected data items. The Graphic Summary illustrates the Nation's agriculture using dot and multi-colour pattern maps. The Coverage Evaluation publication provides estimates of the completeness of the census for the United States and the four geographic regions. The Procedural History presents a comprehensive summary of the planning, preparation, data collection, processing, and publication activities.

## **2. Mail Enumeration in the Census of Agriculture**

### **2.1 History of Enumeration**

The 1969 Census of Agriculture was the first mailout/mailback self-enumerated national census of agriculture. All prior censuses were taken by personal interview in a complete canvass of rural areas. Prior to 1950, an enumerator was given the farm definition and told to obtain questionnaires only for those places qualifying as farms. In 1950 and subsequently, the enumerator was instructed to obtain questionnaires for all places with specified types of agricultural operations. Decisions as to which of these places were

farms were made during the processing of the questionnaires in the central office. This procedure was adopted in an effort to improve the coverage of operations that had accounted for a large portion of the undercounted farms.

In 1954, a history book was introduced for use in each enumeration district to improve coverage. The enumerator was instructed to record the location and identification of every dwelling and of every place with no dwelling but with agricultural operations, provided it was partly or entirely located in the given enumeration district. The enumerators were also required to draw the boundaries of each farm and nonfarm tract on a township sketch form.

Because of the difficulty in finding enumerators in rural areas and the cost of personal enumeration, alternative data collection methodologies were evaluated. Following the 1964 census, a special study was conducted to test the feasibility of a mail data collection using an address list developed from federal income tax returns. A sample of 1964 census farm operator names was matched against tax returns for 1963 and 1964. For each sampled operator, a determination was made as to whether individual income tax Form 1040F had been filed in 1963 or in 1964. The study indicated that on a national basis about 96 percent of farms with total value of product (TVP) of \$2,500 or greater would be included on a mailing list composed of names and addresses from the Internal Revenue Service (IRS). For the farms with TVP of less than \$2,500, about 70 percent of the farms were included on the IRS list. On the basis of this study, a decision was made to adopt the mailout/mailback approach for the 1969 Census of Agriculture using tax return lists of farm operators supplemented by other agricultural lists.

### **2.2 Mail List Development**

In order to implement the mailout/mailback data collection procedures, a mail list development process was initiated.<sup>3</sup> Since complete census results are highly dependent upon a complete mail list, this is one of the most significant phases of the overall task of taking the census of agriculture. The objective

<sup>1</sup> 1982 Census of Agriculture, U.S. Summary and State Data, Vol. 1, Part 51.

<sup>2</sup> 1982 Census of Agriculture, Volume 2, Part 2, Coverage Evaluation.

<sup>3</sup> Dea, Jane Y., Tommy W. Gaulden, D. Dean Prochaska, "Record Linkage for the 1982 Census of Agriculture Mail List Development Using Multiple Sources," 1984 Proceedings of the Section on Survey Research Methods, American Statistical Association.

in the list development process is to compile a complete list that minimizes duplicate records and eliminates nonfarm records. The mail list for the census of agriculture is a prime example of a census list compiled from multiple administrative record sources. The availability and procurement of administrative record files are major requirements for mail data collection in the census of agriculture.

Names and addresses of persons and organizations associated with agriculture are obtained from several primary sources. In 1969, these included the IRS Form 1040F file of farm businesses, the Form 1065 farm partnership file, the Form 1120S small farm corporation file, the Form 943 farm employers file, the Agriculture Stabilization and Conservation Service (ASCS) file, the list frame of the United States Department of Agriculture's Statistical Reporting Service (SRS) in the Northeastern, Southern, and specified North Central States, and an updated large farm list from the previous census. Since 1969, these sources have been expanded to include the files of farm operators from the previous census of agriculture, the nonrespondent file from the past census, the list frame of the SRS for all available states, and special lists from various sources for large or specialized farm operations. However, not all the names on the individual source lists qualify as census farm operations.

The quality of identifier information obtained for the source lists from outside the Census Bureau varies by source. Most of the source lists have a code or value indicating size and type of operation. There is extensive duplication between files and within files. Variations of the same name - nicknames, initials, middle names, and farm names appear in the source lists. Farm operators use different addresses due to business and residential locations or relocations. The nonfarm records and the duplicate records from the previous census are used to aid in the determination of farm status and duplicates.

The development of the census mail list consists of two list building phases: 1) the Farm and Ranch Identification Survey phase (15.8 million source records in 1982), and 2) the mail census phase (an additional 3.2 million source records in 1982). Each phase has five major operational parts: 1) Format

and Standardization; 2) Employer Identification Number (EIN) and Social Security Number (SSN) linkage; 3) Geographic coding and ZIP code edit; 4) Alphabetic name linkage; and 5) Clerical review of all record sets not previously identified as duplicates or nonduplicates. The alphabetic linkage part of the system is based on the record linkage theory developed by Fellegi and Sunter of Statistics Canada.

In 1982, completion of the first phase of record linkage resulted in a file of approximately 7.3 million records. Each of these records was classified either as a probable farm (1.9 million records), as a questionable farm (3.1 million records), or as a probable nonfarm (2.3 million records). The records in the "probable nonfarm" group were removed from the mail list. The records in the "questionable farm" group were selected for inclusion in the Farm and Ranch Identification Survey. The objective of this survey was to identify nonfarm operators and to add new tenant and successor names. Records identified as census farms from this survey, records from previously unavailable source lists, and records classified as "probable farms" were used in the second phase of record linkage to develop the final census mailing list. In 1982, this two-phase process reduced the total source records from 19.0 million total records to a mail list of 3.6 million records.

In 1982, the use of the nonfarm records from the previous census in the linkage operation was effective in reducing the total size and nonfarm composition of the final census mail list. The final 1982 mail list had about 20 percent fewer records than the 1978 list of 4.4 million records. A study of comparability of census data indicated that the final mail list coverage of the 1978 and 1982 censuses was very comparable.<sup>4</sup> The record linkage and development process for the mail list in 1982 cost approximately \$1.5 million. An additional \$4 million was required to conduct the Farm and Ranch Identification Survey.

## **2.3 Limitations of a Mail Enumeration of Farm Operators**

Although the final census mail list is compiled from a large number of source lists, it does not completely cover the universe of census farm operations. Types of operations that often do

<sup>4</sup> Clark, Cynthia Z. F., "Comparability of Data from the Censuses of Agriculture," 1984 *Proceedings of the Section on Survey Research Methods*, American Statistical Association.

not occur on source lists include those where the operator is minimally associated with agriculture or does not identify an operation as a farm. A farm may be erroneously classified as a "probable nonfarm" and eliminated from the list prior to the Farm and Ranch Identification Survey. A farm operation also may be eliminated from the list on the basis of incorrect responses to the screening questions in the survey or because the census response is misclassified as a nonagricultural operation. All of these errors contribute to the "undercount" of census farms. In 1982, we estimated an undercount of 336,000 farms.<sup>5</sup>

The mail list development process introduces another problem that needs to be addressed in census procedures. As previously indicated, many duplicate name and address records appear on several source lists. The record linkage methodology eliminates most of these duplicates. Respondents are encouraged to use special procedures if they receive duplicate forms. However, some duplicates remain. These duplicates plus nonfarms erroneously classified as farms in census processing contributed to an estimated "overcount" of 114,000 farms in 1982.<sup>6</sup>

As previously stated, evaluations of coverage have indicated that the census enumeration is less complete for farms with sales of less than \$2,500. Many of the operators of these smaller farms do not have any formal association with agricultural organizations or even identify themselves as agricultural operators. Because of these characteristics their names are not generally included on source lists used in developing the census mail list. Thus, some other methodology is necessary to provide better coverage of this portion of the farm universe.

In 1978, the census of agriculture conducted an area segment sample in conjunction with the mail enumeration.<sup>7</sup> Data collected from the area sample were used to augment the census at the state and national levels.

This methodology substantially improved the coverage of the census, particularly for farms with sales of less than \$2,500. In 1978, the percent of small farms not on the census was 3.2 compared with 28.5 in 1982 (Table 4).

However, budget restrictions have eliminated the use of this methodology in 1982 and 1987.

Response to any method of data collection has an impact on the quality of the final data. Ensuring that all census recipients complete and return the census is particularly critical for a mail enumeration where the list contains both farm and nonfarm operations. General census procedures described in Section 3.2 have been designed to encourage this response. However, several specific types of agricultural operations have been identified as requiring special procedures to obtain respondent-supplied (rather than imputed) data and review by an agriculture analyst. These large or unique operations, referred to as "musts", bear a special designation in the mail label to facilitate the use of these procedures. "Musts" include multiunits, abnormal farms, and farms with an estimated sales value above a state cutoff (generally \$100,000).

Most of the data items published in the census of agriculture are collected for all mail list respondents. Selected data items, though, are collected for only a sample of mail list addresses. These include data on fertilizer and insecticides, machinery and equipment, expenditures for interest, energy, and production, and value of land and buildings. Because of the impact of large operations on the estimates for these data items, the mail list sample is designed to include with certainty specified addresses from the mail list that were expected to meet size (in acreage or total value of sales) or geographic criteria. The size criteria varied by state from 1,000 to 5,000 acres and from over \$40,000 in sales to over \$200,000. All farms in counties with fewer than 100 farms enumerated in the previous census were also included in the sample. The "must" and "certainty" groups overlapped in that the "must" records that qualified by value of sales were included in the certainty sample stratum.

### 3. Procedures Used in Mail Enumeration

#### 3.1 Publicity and Public Awareness Program

The program of the census of agriculture uses varied avenues for publicity that are targeted

<sup>5</sup> 1982 Census of Agriculture, Volume 2, Part 2, Coverage Evaluation.

<sup>6</sup> *ibid.*

<sup>7</sup> 1978 Census of Agriculture, Volume 5, Part 4, Procedural History.

to reach all levels of the farming sector. Prior to the census, we design posters and information kits for distribution to those willing to help publicize the census. We ask local businesses and farm banks to set up posters in their prominent work areas. Our staff attend trade and farm shows to publicize the census among farmers and to inform users of census data. We distribute pamphlets and one-page leaflets at these shows to emphasize the importance of our report form - "Fill it out - Mail it back."

We solicit help from county agricultural offices such as the County Extension Service, the Agriculture Stabilization and Conservation Service, and the Soil Conservation Service in distributing information, in publicizing the census at agricultural meetings, and in providing assistance in completing census forms. We also prepare a guide providing instructions on how to complete a census form for officials from these offices. In addition, we prepare material directed toward agriculture students for use in vocational agricultural programs. Lesson plans and guides are used to provide information on the importance of the census of agriculture and to motivate the student to help his/her parents fill out the census report form. In past censuses, we have also received favorable publicity from Congressmen, Senators, States assemblymen, and local politicians who have given verbal support to the census of agriculture.

We also solicit the cooperation of broadcasters, and farm newspaper and magazine editors. These media reach into most farm and ranch homes. Television and radio farm broadcasters publicize the census to the farmer and rancher by providing information about the census and census data in conjunction with farm market reports and other farm news. We provide information for short news items and drop-in ads that are placed in local and national farm journals and newspapers. We ask magazine editors to feature census articles, using a cover photograph as a lead-in where possible. We also provide information and advertisements to trade associations for their publications. These publications announce farm shows and events of interest to specialized areas and have a circulation that reads different components of the agriculture universe.

We find all of the above-mentioned methods very useful when there is no direct contact by an enumerator with the respondent. In the initial census mailing, we enclose brochures and pamphlets with the report form explaining why the census is necessary and how the census can be helpful to the respondent. We enclose additional information in followup mailings.

### 3.2 Data Collection

The mailings for the 1969, 1974, 1978, and 1982 censuses used a combination of letters, report forms, and reminder cards. Since the procedures employed were similar in these censuses, the details will be presented for the 1982 census.<sup>8</sup> Census report forms were mailed initially in late December 1982 requesting return by mid-February and followed by six reminders mailed on a flow basis at approximately 3-week intervals. The initial census followup was a postcard reminder mailed after the mid-February due date to all nonrespondent addresses. The second and sixth followup mailings were census report forms with instructional materials. The remaining followup mailings were letters requesting response, pointing out the uses of census data and reminding addressees of their legal requirement to respond to the census. Table 1 provides the date, type, and volume of each mailing and the response rate at the time of the mailing.

At the time of the second mail followup, a file containing names and addresses of non-respondents in the "must" group was selected for later use in telephone followup. This included 417,000 addresses that had a potential farm operation with \$100,000 or more in total sales in 1982. The names on this file were retained on the mail followup file until either a mail or telephone response was obtained. Telephone followup for the "must" group began in May. If neither mail nor telephone response was received for "must" nonrespondents, attempts were made to obtain secondary source information. The 2,700 local offices of the Agriculture Stabilization and Conservation Service (ASCS) of the U.S. Department of Agriculture were the most important source of secondary data. These cases were then edited by an analyst.

<sup>8</sup> Ruggles, Donna R., Jane Y. Dea, Flora K. Kwok, and Cindy A. Carman, "Evaluation of the Effectiveness of Data Collection Procedures for the 1982 Census of Agriculture," 1984 *Proceedings of the Section on Survey Research Methods*, American Statistical Association.

Although the overall response rate to the census had reached 65 percent by April 1983, a number of individual counties had much lower rates. To encourage response from these areas, it was decided to initiate a supplementary followup effort to nonrespondents in selected states or counties with response rates lower than the national level. A special followup letter that offered assistance in completing the census report form was prepared for mailing to this group of 309,000 nonrespondents. In May 1983, a supplementary telephone followup was initiated for a sample of nonrespondents from 252 counties with response rates below 75 percent. The telephone unit that was initially established in Jeffersonville, Indiana to handle incoming calls was expanded to handle the telephone followup operation. Where possible, the unit obtained telephone numbers for nonresponse cases and conducted telephone interviews for those cases. When no response was obtained from "must" cases, the telephone unit attempted to obtain data from secondary sources. The unit also telephoned in-scope respondents who had incomplete data items.

### 3.3 Edit, Item Imputation, and Analytical Review

After a mailed report form was received at the processing site, it went through a check-in procedure.

If the barcode could not be machine-read, the check-in information was keyed. If the form was blank or had correspondence attached, the form was sent to a clerical review section for resolution. The cases were then batched for keying in state order. For the first time with the 1982 census, no clerical pre-edit was performed prior to keying. The keyers were instructed to key the data, with key codes, as reported and to flag obvious respondent alterations, such as wrong units, altered stubs, data-related remarks, bracketed entries, and double entries. Even the written notes and comments on the report form were taken care of by the keyers who keyed special codes. These respondent alterations and coding for "other" crops and livestock were checked in the edit review following the complex edit.

Following the keying, files were transmitted to the Washington Office for formatting. In this process the keyed records were passed through a program that first converted the data into standard units of measure. It then put the data into binary representation and standardized the record format. Each census

file number (CFN) had a fixed length section for identifier information and a variable length section for data values. The data section had computer words for each data item reported, changed, or imputed. Out-of-scope records were identified. Any invalid state, county, and form type codes in the record were identified for correction and rekeying. A correction file was created for incorrect key codes and matched to the main file where the corrected values replaced the errors. The entire record was rekeyed if the number of maximum acceptable rejects was exceeded.

Next, all individual edits and imputations occurred in the complex edit. This operation validated, cross-checked, and refined the reported data. The edit specifications were given to the programmers by subject-matter specialists. Key ratios were tested against tolerance limits based on experience in previous censuses and surveys. The edit routines rounded data items, substituted the sum of detailed items for reported totals, and imputed individual data items from pre-defined ratios.

The computer program for the complex edit probably contained over 10,000 steps, but not all steps were used on every report. The edits were performed separately for each state. During the edit of each census record, farm classification status was determined, missing entries were supplied (item imputation), total crop acreage was reconciled with reported crop acres, reported values outside prespecified limits were changed, values reported for product sales were checked using the average price for the state, other consistency checks were made, and acreage, tenure, TVP, SIC and type of organization categories were determined. In addition the 1982 data were compared to 1978 data. This replaced the previous census procedure requesting analyst review of all "must" cases before keying.

Once the complex computer edit was completed, a printed listing was produced of all data items that failed the edit, data items that were changed by the edit, and flags that were set by the edit. These listings along with the corresponding report form were reviewed by the clerical staff in Jeffersonville. If the clerical procedures were not adequate for resolution, the case was referred to the agriculture analysts. Corrections were prepared and keyed. The correction file was then transmitted to the Washington, D.C. office for matching to the main data file. The corrected files were then re-edited to ensure that the

**TABLE 1. Mail Contact in the 1982 Census of Agriculture**

Mail Contact	Mail Date	Type of Mailing	Number Mailed	Response Rate at Mailing (Percent)	Increase in Response Rate (Percent)
Initial Mailout	Dec. 1982	Letter, Report Form & File Copy, Instructions, "Your Farm or Ranch Counts" Brochure	3,600,000		
	1/28/83			31.4	31.4
First Followup (due date 2/15/83)	2/22/83	Reminder Card	1,900,000	48.4	17.0
Second Followup	3/15/83	Letter, Report Form, Instructions	1,600,000	57.3	8.9
Third Followup	4/13/83	Letter	1,071,000	70.1	12.8
Fourth Followup	5/12/83	Letter	890,000	75.6	5.5
Fifth Followup	5/25/83	Letter	790,000	78.2	2.6
Sixth Followup	6/21/83	Letter, Report Form, Instructions	708,000	80.3	2.1
	7/15/83			83.0	2.7
	9/09/83			85.2	2.2
	Final			85.4	.2

Source: U.S. Bureau of the Census, Data from Agriculture Division Final Mail List Check-in Tabulations.

corrections were properly made and that no further action was necessary.

After the edit and failed-edit corrections were completed, the corrected files for each state were merged into the U.S. detail file in the proper state sequence. At the same time, duplicate records were identified. Unless the records could be identified as different operations under the same CFN, only the first of the duplicate records was retained. The merge program also produced tallies of farms by size, TVP, and type that would be used in the whole farm nonresponse imputation routine.

### 3.4 Nonresponse Imputation

The total census of agriculture data collection effort, consisting of both mail and telephone followup, achieved a response rate of 86 percent. Continuing followup efforts after the scheduled period would result in only a marginal increase in the response rate. In order to publish data for the entire farm universe, collected data were weighted to

account for nonresponding farm operators. A survey of census nonrespondents was designed to provide state estimates of the proportion of in-scope nonrespondents of the total nonrespondents for use in imputing census enumerated data to nonrespondent farms.

The nonresponse sample was a single stage, stratified, systematic sample with selection rates varying by stratum and by state. Nonrespondents within a state were divided into six strata. Strata were based on mail size classification, administrative record source, and special handling codes. Approximately 13,000 names and addresses were selected from the April universe of census nonrespondents. The variable selection rates used for each state were designed to estimate the number of nonrespondent census farms in each state with a relative error of about 6 percent.

The report form for the nonresponse survey was designed to differentiate between addresses of census recipients that qualified as census farms and those that did not.

Sampled nonrespondents were mailed a report form at the end of April, and another report form 2 weeks later. They were next contacted by telephone. Since data collection for the survey was concurrent with the census, if a sampled nonrespondent responded to the census, that nonrespondent was dropped from the sample.

On the basis of the nonresponse survey, the proportion of farm nonrespondents was estimated for each state. A synthetic estimator was then used to estimate the number of nonrespondents by strata in each county of that state. Finally, a sample of respondents was selected to represent the nonrespondents. However, data for a "must" nonrespondent were not imputed using this methodology but rather, as mentioned earlier, obtained either from telephone followup, secondary sources, or historical data.

After all editing and imputation procedures were completed, sample weighting was performed followed by tabulation. Agriculture analysts reviewed the tabulations, prepared detailed criticisms of questionable data, and suggested corrective actions. The clerical staff in Jeffersonville checked the data for duplicate records and for keying, reporting, or processing errors. They obtained additional respondent information where necessary and prepared corrections to individual data records for analyst review. Preliminary reports were prepared and reviewed with corrections being made to the data file as many times as was necessary to ensure record accuracy. An analysis of potential disclosure of individual data was performed prior to preparation of final data tables.

#### **4. Evaluation of Mail Enumeration**

##### **4.1 Quality of Census Published Data**

Publishing quality data obtained from the census of agriculture is complicated because the census mail list contains a large number of addresses (nearly 1.4 million in 1982) that do not qualify as farm operations. In developing the mail list, a number of addresses whose farm status is unknown are retained in order to more adequately cover the farm universe. Because of this, the data collection procedures must be directed to questionable farm operators as well as actual farm operators. The report form must be understandable to both groups in order to obtain response and to ensure that the response is classifiable.

The quality of statistics derived from the census report form is affected by many factors. Among these are: 1) the completeness and accuracy of the mail list of farm operators; 2) the effectiveness of the data collection procedures in eliciting response from the surveyed list; 3) the comprehensibility of the report form and instructions – for this influences the accuracy of respondent supplied information; 4) the accuracy of data processing in correctly classifying response as farms or nonfarms; 5) the completeness of reporting of individual data items by respondents; 6) the extent of the edit and use of secondary source information in single item imputation; 7) the effectiveness of record linkage procedures used in identifying duplicate farm operations; and 8) the reliability of the methods used for estimating data for farm operator nonrespondents.

All of these factors, except the first, are relevant to the possible change in methodology for the 1991 Canadian Census of Agriculture. Evaluation studies of the 1982 and prior censuses of agriculture that relate to factors 1, 2, 4, 7, and 8 are discussed in this section. They include evaluations of response rate and followup procedures (factors 2 and 8), accurate classification of response (factor 4), imputation for nonrespondents (factor 8), and the coverage of the census (factors 1, 4, and 7).

In addition, research studies that are being planned in conjunction with the test of the 1987 Census of Agriculture will provide more insight for factors 2, 3, and 5. These are discussed in Section 5, Initiatives for the Future. Research is being conducted in the next few years that will provide more information on edit and item imputation (factor 6) and on nonresponse imputation (factor 8). The coverage evaluation program for the 1987 Census of Agriculture will provide additional insight for factors 1, 4, and 7 as it will reflect any methodological changes from previous censuses in mail list development, record linkage, and classification procedures.

##### **4.2 Response Rates as a Measure of Data Quality Effectiveness**

The response rate for a survey is a standard measure of the effectiveness of the data collection in eliciting response from the surveyed universe. Examining various aspects of census response over time provides several different insights into the effectiveness of the agricultural census data collections. Published census response rates are calculated as the quotient of all receipts (including forms

returned by the post office - Post Master Returns or PMRs) divided by the total number of addresses on the mail list. On this basis the response rate of 88.0 percent for the 1978 census is considerably higher than that of 85.4 percent for the 1982 census, and somewhat higher than the 1974 rate of 87.4 percent (Table 2). Since there was a proportionately larger number of PMRs in 1978 than in 1974 or 1982, this definition of response somewhat overstates the effectiveness of the 1978 data collection effort in relation to 1974 and 1982. Removing the PMRs from receipts and from the total mail list gives response rates of 85.1 percent, 87.3 percent, and 87.1 percent for the 1982, 1978, and 1974 censuses respectively. On this basis the response rate in 1982 was 2.1 percent lower than in 1978.<sup>9</sup>

The broad universe covered by the farm definition complicates collection of the desired data. The initial report form and accompanying letter, and the mail followups may not effectively communicate to all recipients the necessity for their response, whether or not they perceive that their activities are agricultural. Thus, a number of farm and non-farm operations are not reported. Telephone interviewers may be more effective in obtaining information leading to identification of farm or nonfarm status from nonrespondents who do not perceive that their activities are agricultural. This procedure, however, was primarily used in 1982 to obtain information from nonrespondents who were thought to have either a large or a unique farm operation.

The final 1982 Census of Agriculture data were based on 3.1 million responses from a mail list of 3.6 million names and addresses. Of these respondents, 67.4 percent were agricultural operations. There are several procedural factors that might have affected the response rate for the census of agriculture. These include differences in the response rate by type and by frequency of followup. Another factor thought to affect response is the length of the form. Response rates over time as well as by census mail list classification of size (measure of size derived from indicators present in mail list source records) were examined to gain some insight regarding the optimum frequency of followup mailings.<sup>10</sup>

Weekly response rates for the period January 21 through September 9, 1983, were calculated as the total number of returns divided by

the total number of report forms mailed out. Returns consisted of all report forms mailed back (whether completed or not), all correspondence from the report form recipients, and undeliverable report forms returned by the post office. All potential farms to whom report forms were mailed initially were classified into 16 categories based on their expected 1982 sales. These categories then were aggregated into five groups for this study. The expected sales of these groups were: A—at least \$100,000, B—\$10,000 to \$99,999, C—\$1,000 to \$9,999, D—less than \$1,000, and E—unknown.

About 25 percent of the mail list addresses were mailed long report forms. The long report forms contained all of the questions that were on the short report forms as well as some additional questions. The recipients of long report forms came from two groups. The "certainty" group consisted of recipients whose size and source code indicated a "large" (expected sales of \$40,000 or more) farm operation. Approximately 328,000 or 9 percent of the mail list were certainty cases. The other recipients of long forms were sampled from the remaining addresses on the mail list. These 573,000 recipients of long forms were referred to as the "uncertainty sample." The remaining mail list addresses, 2.8 million, were mailed short forms.

A cumulative national response rate of 46 percent was achieved by the February 15 due date. The cumulative national response rate increased at the highest rate between January and the middle of April, tapered off until mid-July, and levelled off from then until the end of the data collection period (Figure 1). Each of the six mail followups to nonrespondents was effective in increasing the response rate (Table 1). Of these mail followups, only the second and the sixth contained a report form with the letter. The largest weekly increase occurred three weeks after the second followup with a 12.8 percent increase between the second and third followups (Figure 2). Three weeks after the sixth followup, the response rate was higher than in the ten preceding weeks with a 2.7 percent increase. This implies that a report form may be more effective than a letter in eliciting response.

The response rates for all of the potential farm size groups, except for those whose size was

<sup>9</sup> Clark, Cynthia Z. F., op. cit.

<sup>10</sup> Ruggles, Donna R., et al., op. cit.

**TABLE 2. Census of Agriculture Mail List Response**

	1982	1978	1974
Mail List Size	3,654,674	4,429,633	4,182,374
Post Master Returns (PMRs)	82,792	230,980	108,700
Mail List - Excluding PMRs	3,571,702	4,198,653	4,073,674
Nonresponse (includes remails)	531,916	532,030	525,875
Receipts	3,039,966	3,666,623	3,547,799
In-scope	2,021,400	2,044,989	2,029,389
Out-of-scope	978,264	1,511,218	1,487,351
Nonclassified	40,302	110,416	31,059
<b>TOTAL MAIL LIST</b>			
% Overall Response Rate	85.4	88.0	87.4
% Classified Respondents	82.1	80.3	84.1
% Other Respondents (PMRs, Nonclassified)	3.3	7.7	3.3
% Nonrespondents	14.6	12.0	12.6
<b>MAIL LIST EXCLUDING POST MASTER RETURNS</b>			
% Overall Response Rate	85.1	87.3	87.1
% Classified Respondents	84.4	84.7	86.3
% Nonclassified	1.1	2.6	.8
% Nonrespondents	14.9	12.7	12.9
Source: U.S. Bureau of the Census, Data from Agriculture Division Final Mail List Check-in Tabulations.			

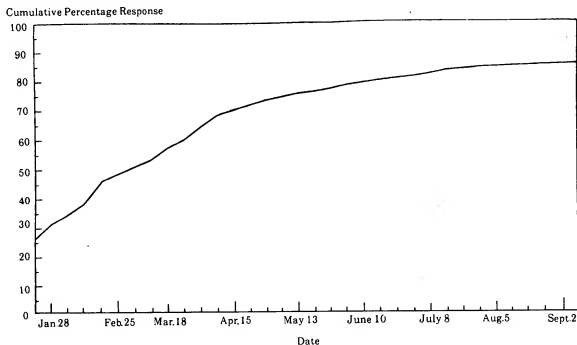
unknown, followed a similar pattern of increase between January and the middle of June (Figure 3). The mail followups had about the same effect on each size group. The telephone followup to all nonrespondents in size group A (beginning in mid-May) caused the cumulative response rate for that group to increase at a faster rate than the cumulative response rates for the other groups. By the end of June, group A had the highest cumulative response rate among all groups and the response rate of A continued to increase until data collection ceased, achieving a final response rate of 97.7 percent. This relatively high response rate for group A illustrates the effectiveness of the telephone followup operation.

The cumulative response rates for groups B, C, and D had similar patterns of increase

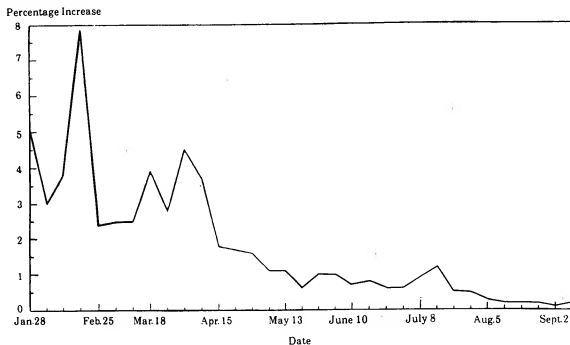
between the middle of June and the end of the tabulation period. The final response rate for group E was substantially lower than for those addresses for which an expected size classification was possible. The proportion of actual farm operations among respondents in this group was considerably lower than among the other groups reflect. The low response rate for this group may reflect inadequate instructions as to who should complete the census form.

The length of the report form did not seem to have much impact on the response rates (Figure 4). However, in June, telephone followup began for the certainty portion of the group receiving the long form, thus increasing the response rate for that portion of the long form group. Response rates for the sample portion of the group receiving the long form

**Figure 1. Cumulative National Response Rate**



**Figure 2. Weekly Increase in Cumulative National Response Rate**



and the nonsample group receiving the short form were very close. This indicates that the additional respondent burden associated with the longer report form did not have an adverse effect on response.

On the basis of this evaluation of response rates, several conclusions can be reached. The mail followups in which a report form was sent along with a letter appear to be more effective in increasing the response rates than the mail followups in which only a letter was sent. This idea was explicitly tested in a mail variation test. However, the telephone follow-up to all nonrespondents whose expected sales were \$100,000 or greater was the most effective procedure for increasing the response rate for that group. The length of the report form did not appear to affect response rate to the census. The response of census recipients grouped by expected sales did not appear to be correlated with response rate.

#### 4.3 Effectiveness of Varying Followup Procedures

The objective of the mail variation test was to determine if there was a statistical difference in mail response between a report form followup and a letter followup.<sup>11</sup> In order to test this hypothesis, a test group and a control group were selected. The procedure used for the initial mailout and first followup for the test group and control group was identical to that used for the census. For the second and third followup mailings, the test group received first a letter and then a report form, reversing the order of followup used in the census and for the control group.

Cost considerations limited the sample selection to 13 states and a sample size of 100,000. The states from which the sample was drawn were chosen because they were representative of two very different areas in terms of farm size. Seven of the states (Virginia, North Carolina, South Carolina, Georgia, Kentucky, Tennessee, and Alabama) were from the South where there are more small farms (in both size and product value), and for which the state response rates have been the lowest in past censuses. The remaining six states (Ohio, Indiana, Illinois, Iowa, Nebraska, and Kansas) were selected from the Midwest where the response rates have been among the highest for past censuses. An additional factor in the design of the experiment was the length of the report

form - long and short. The systematic sampling procedure selected approximately 5,000 mail list addresses per report form and group from each of the 13 states.

Up until the first time point (March 19), the test group and control group had been treated identically by receiving the initial report form and a postcard reminder. There was no significant difference in response at this point. March 19 was during the mailout of the second followup in which the control group received a report form and the test group received a letter. A second time point (April 23) was chosen at the end of the mailout of the third followup. By this time both the test group and control group had been mailed a letter and a report form, but in different order. The final test point (May 21) occurred during the fourth followup mailing.

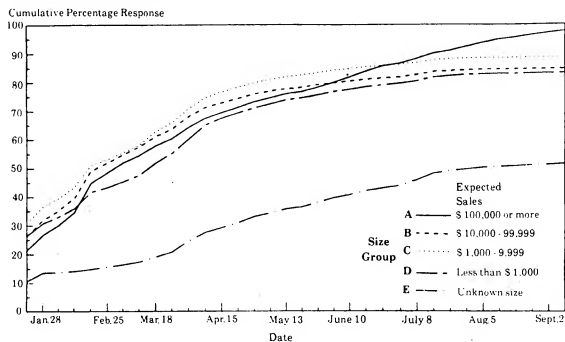
A multivariate analysis of covariance model was used with the cumulative response rate at time one as the covariate and the cumulative response rate at time two and three as the dependent variables. Three factors: group (control versus test), report form (long versus short), region (South versus Midwest), and one interaction term (form by region) were represented in the model. The analysis showed that there was a group difference and a region difference, but no significant form or interaction difference. Also, the covariate had a mean difference between regions and was a significant term in the model. Figure 5 visually suggests the results of the analysis.

The analysis of covariance model with the same terms as above also was used in a second analysis. In this model, time two (April 23) was used as the covariate and time three (May 21) was used as the dependent variable. The results were the same as the first analysis - a group and region difference but no significant form or interaction difference.

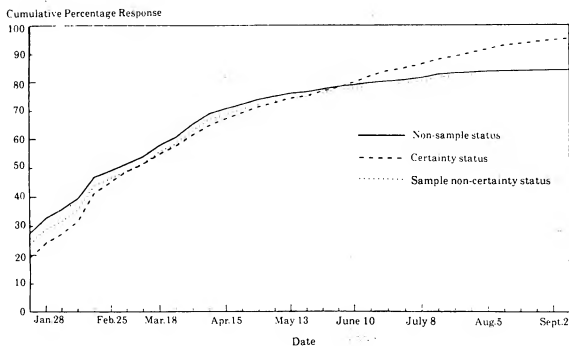
The results of the mailout variation test clearly indicate that report form followups are more effective than letter followups (Table 3). This conclusion will be important in planning the program of mail followup for the 1987 Census of Agriculture. A key step toward improving this program will be to conduct a cost analysis of the more costly report form followups versus the less costly letter followups. An important factor in evaluating the cost-effectiveness of the report form followup will be the potential increase in data quality obtained from earlier response.

<sup>11</sup> Ruggles, Donna R., et al., op. cit.

**Figure 3. Cumulative Response Rates by Farm Size**



**Figure 4. Cumulative Response Rates by Sample Status**



#### 4.4 Evaluation of Telephone Followup

As part of the evaluation of telephone follow-up procedures for the "must" nonrespondents, two samples of 10,000 "must" nonrespondents were selected when the nonrespondent file was created. The samples were selected by means of a stratified cluster sample within each state where the sample size for each state was determined by the proportion of eligible nonrespondents in that state of the total. Within each stratum in a state a systematic sample of pairs of nonrespondents was selected with cases within each pair randomly assigned to one of the two samples. One sample was interviewed using the regular telephone methods at the Census Bureau's Data Preparation Division located in Jeffersonville, Indiana. The other sample was interviewed using computer-assisted telephone interviewing methods (CATI) at the Census Bureau's Washington, D.C. office.<sup>12</sup> A special CATI version of the census of agriculture questionnaire was developed from the questionnaire designed for telephone interviewing.

Analysis of the results from the two methodologies will include distribution of final resolution of sample cases, rates for each type of resolution, average number of calls per case to reach a final resolution, average total minutes per case to reach final resolution, and percent of item changes in the complex edit. Preliminary results indicate that CATI response rates were slightly higher than those obtained in regular telephone interviewing and that the quality of the data resulting from the two methodologies was about the same. A paper detailing the results will be available by the end of 1985.<sup>13</sup>

#### 4.5 Accurate Classification of Responses

Responses from farmers and nonfarmers, responses not classified at the time of the closeout of the census data collection process, and forms that are returned by the post office have historically been included as receipts in calculating final response rates for the census (Table 2). A comparison measure of

classification of response from the past three censuses at the time of closeout can be derived from a breakout of these receipts.<sup>14</sup> Excluding the post master returned forms from the total number of mail list addresses, the percent of forms that were classified at closeout was very comparable for the 1982 and 1978 censuses (84.4 percent versus 84.7 percent). This was achieved because the 1982 census processing system classified a larger proportion of the census receipts (excluding Post Master Returns) by closeout than the 1978 census.

Once responses are received from mail list recipients they must be classified. The tabulations discussed in the previous paragraph gave a measure of responses that had not been classified at the end of the data collection period. The coverage evaluation conducted for the census has provided two measures of error in classification - a measure of actual farms classified as nonfarms (misclassified), and a measure of nonfarms classified as farms that includes multiple counted operations (overcount). Percentages obtained by dividing these measures of classification error by the estimated farm universe can be compared (Table 4). In the category of farms whose total sales were \$2,500 or greater, misclassified farms decreased over these censuses. However, the estimated percent of overcounted farms was greater in 1982 than in previous censuses.

#### 4.6 Imputation in the Census

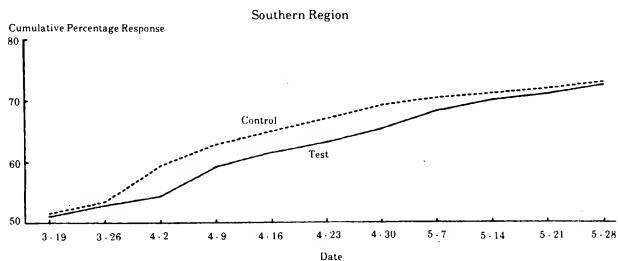
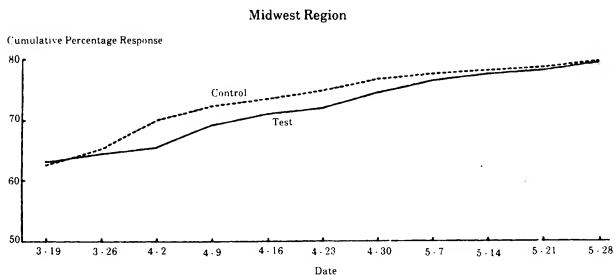
Because there are many addresses on the mail list that do not represent farm operations, not all nonrespondent addresses represent farms. The agricultural census data estimating procedures adjust for farm nonrespondents by estimating the proportion of nonrespondents on the mail list that are farm operators for each state, imputing values for data items for that number of nonrespondents, and incorporating the imputed data into the estimation procedure for each published data item. As information obtained from the respondent is generally believed to be more accurate than imputed data, the percent of the published data that are respondent supplied gives a

<sup>12</sup> Ferrari, Pamela, Richard R. Storm, and Francis D. Tolson, "Computer-Assisted Telephone Interviewing in the 1982 Census of Agriculture," 1984 Proceedings of the Section on Survey Research Methods, American Statistical Association.

<sup>13</sup> Ferrari, Pamela W., "An Evaluation of Computer-Assisted Telephone Interviewing Used During the 1982 Census of Agriculture," U.S. Census Bureau Internal Report.

<sup>14</sup> Clark, Cynthia Z. F., op. cit.

**Figure 5. Cumulative Response by Region in Mail Variation Test**



**TABLE 3. Percent Imputation in Census of Agriculture Data**

	Mail List 1982	Mail List Only 1978	Mail List & Area Sample 1978	Mail List 1974
<b>PUBLISHED FARMS</b>				
% Mail List	90.2	82.5	90.7	87.7
% Area Sample	NA	8.9	NA	NA
% Imputed	9.8	8.6	9.3	12.3
<b>LAND IN FARMS</b>				
% Respondent Supplied	94.4	95.4	95.3	94.1
% Imputed	4.6	4.6	4.7	5.9
<b>HARVESTED CROPLAND</b>				
% Respondent Supplied	94.1	93.5	93.4	93.8
% Imputed	5.9	6.5	6.6	6.2
<b>VALUE OF AGRICULTURAL PRODUCTS SOLD</b>				
% Respondent Supplied	96.3	96.1	96.1	95.9
% Imputed	3.7	3.9	3.9	4.1

Source: U.S. Bureau of the Census, Agriculture Division.

measure of data quality.<sup>15</sup> The percent of imputation of data for an entire farm operation was between 9 and 10 percent (Table 5) for the two recent censuses, but was 12.3 percent for 1974. As previously noted, although the response rate was higher in 1974, the proportion of published farm operations with imputed data was higher than in either 1978 or 1982.

The proportion of respondent supplied data for other major data items – land in farms, harvested cropland, and value of agricultural products sold—has consistently been higher than for the published farm count. Because the census of agriculture has a more intensive followup procedure for mail list nonrespondents whose expected sales are large, most of the farm nonrespondents for which data are imputed have small farm operations. Due to the small size of these operations, these data have less impact on the values of land in farms, harvested cropland, and product sales value. The estimation properties of the imputation methodology used for census farm nonrespondents is also a factor in the quality of the published estimates. A study of

alternative imputation methods for entire farm operations is planned prior to 1987.

#### 4.7 Coverage of the Farm Universe

In order to provide an independent measure of the number of farms not accounted for in census published data, a coverage evaluation program has been conducted for the census of agriculture since 1945.<sup>16</sup> The 1978 and 1982 coverage evaluation samples were designed to provide regional level estimates of several components of census coverage rather than the state level estimates provided in 1974.

Estimates of total farms in the universe are provided in the coverage evaluation publication where "estimated farms" is the sum of farms "on the census", "not on the census", and "misclassified on the census" minus the farms "overcounted" in the census. Each of these estimates was calculated<sup>17</sup> for three categories of farms – all farms, whose total sales are under \$2,500 (small farms), and farms whose total sales are \$2,500 or greater. The estimate of these components for the three censuses under consideration is given in

<sup>15</sup> Clark, Cynthia Z. F., *op. cit.*

<sup>16</sup> Davie, William C., Emily Lorenzen, D. Dean Prochaska, "Coverage Evaluation for the 1982 Census of Agriculture," 1984 *Proceedings of the Section on Survey Research Methods*, American Statistical Association.

<sup>17</sup> Clark, Cynthia Z. F., *op. cit.*

TABLE 4. Coverage Evaluation Estimates for the Census of Agriculture

COVERAGE ESTIMATES									
UNITED STATES									
All Farms	1982 Mail List	1978* Mail List	1978** ML + AS	1974 Mail List	UNITED STATES All Farms				
Census	2,240,964	2,257,775	2,475,612	2,314,013	Census	91.0	91.4	97.0	91.1
Not on Census	259,944	169,997	43,668	266,495	Not on Census	10.6	6.9	1.7	10.5
Misclassified	76,554	58,144	57,408		Misclassified	3.1	2.4	2.2	
Overcounted	113,623	15,456	23,962	41,026	Overcounted	4.6	0.6	0.9	1.6
Universe Estimate	2,463,839	2,470,460	2,555,722	2,539,489	Universe Estimate	100.0	100.0	100.0	100.0
Net Coverage	222,875	212,685	77,110	225,469	Net Coverage	9.0	5.6	3.0	8.9
US					US				
TVP \$2,500					TVP \$2,500				
	1982 Mail List	1978 Mail List	1978 ML + AS	1974 Mail List		1982 Mail List	1978 Mail List	1978 ML + AS	1974 Mail List
Census	536,327	460,535	611,623	616,728	Census	71.4	80.7	94.6	79.8
Not on Census	213,564	94,588	20,562	164,213	Not on Census	28.5	16.6	3.2	21.3
Misclassified	51,119	17,664	17,664		Misclassified	6.8	3.1	2.7	
Overcounted	50,624	2,208	3,048	8,482	Overcounted	6.7	0.4	0.5	1.1
Universe Estimate	750,686	570,579	646,801	772,459	Universe Estimate	100.0	100.0	100.0	100.0
Net Coverage	214,359	110,044	35,178	155,731	Net Coverage	28.6	19.3	5.4	20.2
US					US				
TVP \$2,500					TVP \$2,500				
	1982 Mail List	1978 Mail List	1978 ML + AS	1974 Mail List		1982 Mail List	1978 Mail List	1978 ML + AS	1974 Mail List
Census	1,702,973	1,794,938	1,864,687	1,695,047	Census	99.5	94.6	97.8	96.0
Not on Census	76,080	75,409	23,102	102,282	Not on Census	2.7	4.0	1.2	5.8
Misclassified	25,435	40,480	39,744		Misclassified	1.5	2.1	2.1	
Overcounted	62,999	13,248	20,914	32,544	Overcounted	3.7	0.7	1.1	1.8
Universe Estimate	1,711,489	1,897,579	1,906,619	1,764,785	Universe Estimate	100.0	100.0	100.0	100.0
Net Coverage	8,516	102,641	41,932	69,738	Net Coverage	0.5	5.4	2.2	4.0
US					US				
Abnormal Farms	1,664	2,302	2,302	2,238					

\* 1978 Mail List; 1978 estimates derived from the census.

\*\* 1978 ML and AS; 1978 estimates derived from the census mail list and the area sample.

Source: U.S. Bureau of the Census, Agriculture Division, Coverage Evaluation Estimates.

Table 4. For comparison purposes with the 1982 and 1974 coverage estimates, separate coverage evaluation estimates have been calculated for the 1978 mail list data. The 1978 published coverage evaluation estimates were designed to measure the coverage of the dual-frame census (mail list augmented by the area frame).

Table 4 presents each of the coverage components as a percent of the coverage estimates of the census total. Over this time period the coverage sample estimate of the percent of farms "on the census" for all farms and for small farms was higher for the 1978 dual-frame census. However, for farms whose total sales were \$2,500 or greater, the 1982 coverage sample estimate of percent of farms "on the census" was higher than the 1978 dual-frame estimate. The percent estimate of farms "not on the census" (not on the mail list or the area sample in 1978) was much lower for all classes of farms for the 1978 dual-frame than for the other censuses. The classification error estimates – both farms classified as non-farms (misclassified) and nonfarms classified as farms (overcounted) – were higher in 1982 than in 1978 for all classes except misclassified farms with sales of \$2,500 or more.

A measure for comparing the relative impact of these components of coverage is the net coverage – the number of farms "not on the census" plus the number of farms "misclassified" minus the number of farms "overcounted." The percent net coverage for the 1978 dual-frame estimates was lower for the category of all farms (3.0 percent) and much lower for small farms (5.4 percent). However, because of a relatively larger estimate of overcount in 1982, the corresponding percent net coverage for farms whose total sales were \$2,500 or greater was less in 1982 than for all previous censuses (0.5 percent).

## 5. Initiatives for the Future

### 5.1 Precensus Test

A precensus test provides a unique opportunity to evaluate alternative methodology and procedures that affect the census program. These include, but are not limited to, report form design (i.e. color, format, type size, etc.), question wording, use of supplementary materials, mail contact and followup procedures, response burden, and attitudes affecting response behaviour. A test of the U.S. Census of Agriculture will be conducted in early 1986. The specific objectives of this test include:

1. test the effect on response of two alternative report formats – a booklet form and a fold-out form similar to the type used in 1982;
2. estimate response burden involved in completing the census report form;
3. test the effectiveness of different types of mail contact and followup;
4. test alternative wording and format of 1982 questions and of any proposed new questions and obtain reasons for item nonresponse;
5. measure quality and completeness of item reporting;
6. test an expansion of the report form sort that occurs prior to keying;
7. test an alternative data keying procedure;
8. evaluate changes in the computer edit routines.

A combination of several methods of testing will be used to achieve these objectives. Approximately 45,000 forms will be mailed in late December of 1985 with a due date of February 1, 1986. Several alternative mail followup procedures will be tested during the data collection period extending through the first 4 months of 1986. Interviews will be conducted in several distinct cluster counties in March. The primary focus of these interviews will be to reinterview respondents to obtain information on the quality and completeness of reporting. Also during March, classroom experiments and observation studies will be conducted in the same general geographic area as the cluster counties. The data processing tests will be conducted in the Jeffersonville, Indiana office that receives the completed forms. A sample of the forms will be keyed to permit measuring item response rates and evaluating changes in the computer edits of the report form.

The mail sample of the census has been designed to provide a statistical test of the effect on response of form type using a booklet and a fold-out report form as measured by overall response rate. Different sessions of classroom observation groups will be used to provide information on perceived burden of completing the two types of forms. Estimates of the response burden of completing the report form will be obtained from several

**TABLE 5. Test of Mail Contact for the 1987 Census of Agriculture**

Mail Date	Followup Received Date	Followup Plan A Type	Followup Plan B Type	Followup Plan C Type	Expected Plan D Type	Response Rate
12/01	12/15	-	-	Precensus Card	Precensus Card	0
12/15	01/01	Initial Mailout	Initial Mailout	Initial Mailout	Initial Mailout	0
01/10	01/22	Thank you/Reminder Card	-	-	Thank you/Reminder Card	0
02/05	02/10	Report Form	Report Form	Report Form	Report Form	50
03/07	03/12	Card/Letter	Card/Letter	Card/Letter	Card/Letter	68
03/29	04/03	Report Form	Report Form	Report Form	Report Form	75

sources. The actual time required to complete the form will be measured in the classroom experiments and observation studies. This, however, will not include time necessary to gather information to complete the form. The estimated time required to complete the form also will be obtained from the respondent's recall on the reinterview. These time estimates will permit comparisons of response burden for the two formats.

Several different plans for a sequence of mail contacts will be evaluated in the test. The plans vary the type and number of mail contacts but not the time between contacts. The types of mail contact used in the plan include a precensus letter accompanied by a card providing census statistics, a thank-you reminder card, followup letters with and without copies of the report form, and followup cards. The sample is being designed to provide a separate test of a group of records that are less likely to be census farm operations. The sample size is sufficient to detect a difference of 2 percent in the response rates for the plans with a .05 level of significance.

The classroom experiments and observation studies will also be used to test alternative wording and format of 1982 questions and of proposed new questions. The classroom experiments would permit randomized distribution of several alternatives. The observation studies using one version at a time would provide insights into response and data collection problems with individual items. The quality and completeness of reporting will

be measured by item response rates from the mail test as well as from the classroom experiments. The quality and completeness of reporting can also be measured for selected items on the reinterview.

Two alternative report form sort procedures that require variation in the number of forms keyed will be evaluated using several time tests at the Jeffersonville, Indiana data processing center. The number of keystrokes and the keying error rate will be compared using forms from eleven different geographic areas for two different data keying procedures. The keyed data will permit evaluation of changes in the computer edit routines.

## 5.2 Response Research

Achieving a high response rate for the census of agriculture is a key component in improving the quality of the statistics resulting from the census. Maintaining good response from a mail data collection is a problem that needs to be continually addressed. Two years ago, a research project examining factors affecting response to the census of agriculture was begun. The research project, thus far, has identified some census procedures that have an impact upon response to the census and has proposed studies that will provide information for changing census procedures to improve response. The objective of the response research is to increase the response rate of all census recipients - both farm and nonfarm. This will reduce the proportion of the mail list for which data imputation is needed, and

consequently improve the quality of the census data.

The content and design of the census mailings could affect the level of response to the census. In preparing for the census test, all the mailed material has been reviewed. The census mailings have primarily been directed to an audience of farm operators. However, a large number of census recipients are not or do not think of themselves as farm operators. Responses to the census though are as important from nonfarmers as from farmers. We have tried to address this problem in developing the letters, cards, and instructional material for the test.

The publicity for the census can be designed to address particular response problems. We are presently evaluating response patterns over the entire data collection period for different geographic areas. Criteria are being developed for low response counties. We will examine farm operator and demographic characteristics in those designated counties. We will then develop publicity designed to address these response needs.

We also are evaluating response patterns of different mail list sources and of different expected sales values. The yield of farm addresses and the response rate of various mail list sources differ considerably. Criteria will be developed for low response sources and sizes. Then differential mail and telephone followup procedures will be designed for nonrespondents in these categories that supplement the general mail followup plans being tested.

A study to examine specific reasons for nonresponse to the agriculture census is being designed. Such a study is done best in conjunction with the census in order to obtain responses from both respondents and nonrespondents that are not affected by a time lapse. A telephone survey for this purpose is currently being planned in conjunction with the 1987 Census of Agriculture. In addition to examining specific reasons for agricultural census nonresponse, it will be designed to provide projected estimates of the response rate and to indicate when alternative census procedures should be used.

# A PROPOSAL FOR A LAND-BASED CENSUS OF AGRICULTURE

G. OLIVER CODE

AGRICULTURE AND NATURAL RESOURCES DIVISION  
STATISTICS CANADA

## Introduction

The purpose of this paper is to outline the concept of a Canadian land-based Census of Agriculture compared with the current headquarters-based census. The proposal pertains primarily to the crops or land-use parts of the census with particular reference to the needs of the Crops Section of Statistics Canada. Based on the advantages of the land-based concept, it is recommended that it be targeted as the eventual desired approach, following a dual-track approach for the 1991 Census of Agriculture. A recommendation is also made for a new procedure for gathering the crop data, and the benefits and problems of the new approach are reviewed. The operational research necessary to implement this type of change in time for the 1991 Census is outlined in the final section.

## The Headquarters Rule

The headquarters rule states that all data on the census document are assigned to the geographic location of the tract of land designated as the headquarters of the farming operations by the farm operator. This implies that any part of the operation located outside that geographic unit will be treated as if it fell within the boundaries of the headquarters unit.

## Limitations of the Headquarters Rule

This is a very practical and simple rule obviously designed to simplify data collection and compilation. Yet the application of the rule may lead to geographic mislocation of particular parcels of land. Historically, this location distortion was generally minor due to smaller and more consolidated farms. In more recent years the adoption of more dynamic and diverse land ownership and leasing combined with larger farm sizes composed of non-contiguous parcels of land has substantially increased the distortion potential.

The Census of Agriculture staff conducted a case study investigating the impact of geographic distortion due to the headquarters rule on the 1981 Agricultural census data for Saskatchewan Census Consolidated Subdivision 410. It was concluded that the distortion measured for a typical Census Consolidated Subdivision was significant.<sup>1</sup> The distortion at this level was at least +10% of the reported land and as high as -44% at the enumeration area level. The Enumeration Area is that area enumerated by one Census Representative during the census. The boundaries of an enumeration area must not cross municipal boundaries or federal electoral district boundaries. A Census Consolidated Subdivision is an aggregation of enumeration areas and is generally the smallest, most commonly used level of tabulation available from the Census of Agriculture.

In an effort to determine the extent of the problem, the Census of Agriculture staff recently compared the total farm area reported in the 1981 Census to the total physical land area at the Census Consolidated Subdivision level. The study focused on the Prairie region of Canada which contains more than 80% of total Canadian improved agricultural land. It was found that 19% of the Prairie Census Consolidated Subdivisions had more farm land allocated through the headquarters rule than the total possible land area. Saskatchewan was the highest at 24%. These percentages are an indication of the seriousness of the distortion. More research is needed to better define the full extent of the problem.

The average farm size in Saskatchewan has increased by almost 15 acres per year since 1966. This established trend to larger farms with apparently more dispersed holdings is continuing and logically, therefore, the potential for distortion is also increasing. Complex leasing, renting and partnership arrangements along with non-resident farm operators<sup>2</sup> present difficult and

<sup>1</sup> Burroughs, R. J., "Investigation into the Impact of Geographic Distortion Due to the Headquarters Rule", Ottawa, Statistics Canada, February 1983.

<sup>2</sup> A non-resident farm operator: can be defined as a farm operator who does not reside on any part of his agricultural holding.

often confusing coverage situations such as overlap and under-reporting. Also there are some cases where the definition of location of headquarters is a problem. As long as the Census of Agriculture uses the headquarters rule, these types of problems can never be completely resolved.

The Census of Agriculture staff have acknowledged the location distortion problem and have some procedures in place to resolve large and obvious land allocation problems. The census staff contact individually the largest corporate farms in Canada in order to correctly allocate their land; the records of non-resident farmers are verified and an examination is conducted of any reported farm land in large urban areas. Despite this, there are Census Consolidated Subdivisions with serious land base data inconsistencies.

The users are demanding more dependability and flexibility at the small area level. The demand is increasing for more data to address economic, environmental, resource, farm management and production issues. The provincial agricultural statisticians have said that the accuracy and coverage of the census small area data are inadequate for many of these tasks or more seriously have errors which may support incorrect analytical conclusions.

The Crops Section of the Agriculture and Natural Resources Division produces a series on acreage, yield and production of the major grains. This crop-reporting system uses small area data at the crop district level for weighting and sample selection. An accurate area frame is considered absolutely essential to the generation of good probability estimates.

### **Recommended Method – The Land-based Approach**

The Census of Agriculture should move toward a land-based census for its crop and land use items. The spatial identity and the related use of the land would be maintained throughout the census system. The reporting unit would be an area location and the data collection would have to record the content of spatial-defined tracts of land instead of farms.

### **Benefits of a Land-based Census**

Compared with the increasing disadvantages of the headquarters rule, the land-based approach offers a variety of advantages since the actual location of the crop data would be maintained.

The crop reporting program of Statistics Canada would benefit substantially from this type of data. With a complete sample frame we could return to the concept of using an area frame, sampling land, and get away from the list frame sampling of farmers. This would mesh well with our remote sensing program which can only employ land-based information. With these land-based data it becomes more operationally feasible to use remote sensing information to generate acreage data and possibly even yields thus reducing the questionnaire response burden on farm respondents. Given the potential developments in satellite technology, the year-to-year acreage changes could then be monitored by satellite and used to generate good small area intercensal data. Potentially, by 1996, satellite remote sensing could be a major component of the collection operations for the crops portion of the census.

The overall coverage of the census would be improved by using easier to define land-based census procedures. For example, the under-reporting and overlap coverage problems would become readily apparent and easier to resolve. The size of farm, or complex leasing, or renting arrangements would not matter because identification of the operator of the land becomes almost unimportant.

The uses of a land-based census data bank would be varied. They range from environmental issues to land use issues. The opportunities for cost recovery should increase dramatically since we will be able to service new and different requests. For example, the Lands Directorate of Environment Canada has a land-use monitoring program that is updated every five years. The census could provide more input to this project.

There exist several administrative data files that are land-based (i.e. soil type maps, crop insurance data, Canadian Wheat Board permit books). These files could be used to link with a land-based census to derive completely new data sets and problem-specific information series.

### **Implementing the Land-based Approach – Some Alternatives**

There are a variety of methods that can be utilized in a move towards a land-based Census of Agriculture. Among these are:

- (a) The United States' Census of Agriculture asks "In what county do you expect the largest value of your agricultural products to be raised or produced?". The U.S. system simply

uses sales instead of headquarters location to allocate land; however, this results in the same types of problems as the Canadian system.

- (b) A more rigorous application of the existing agricultural census data re-allocation methods used by Statistics Canada could reduce the problem. These procedures, as outlined under the section on limitations, however, do not offer a solution or assist in the small area-analysis problems.
- (c) The spatial location of the individual parcels of land in the farm holding could be included in the data capture system. This would allow computer identification of problem areas and some re-allocation of land. Once again this would be a problem-reduction technique, not a long-term solution.
- (d) A matrix questionnaire could be developed in order to allocate a unique identifier to each individual land use and its location on the census questionnaire. This method would allow compilation of a large geocoded data bank that could be readily re-formatted to meet almost any selection criterion. The limitation here is that large matrix questionnaires are notoriously difficult to complete accurately.

### **Recommended Data Collection Technique**

One of the strengths of the Census of Agriculture has been the consistent nature of the questions and collection techniques over time. It is for this important reason that a dual system of data collection is being recommended for 1991.

The existing collection system, based on the headquarters rule, should be maintained for all the current census variables including crops. The enumerator would supplement these data by asking the farmer to outline the location and identify the type of the different crops. It is too early to determine the best and least expensive data recording system but it would probably be either an aerial photograph, a topographic map or a satellite image. The data could then be digitized and used to produce a geographic data base.

### **The Positive Features of the "Mapping-oriented" Approach Are:**

- (a) accuracy of the current census system will be improved due to the necessity to account for all the land in a given area on a map or photograph;
- (b) farmers appear to readily accept and even enjoy studying aerial photographs, etc. As a result, there will be little or no perceived additional response burden;
- (c) the Agriculture and Natural Resources Division of Statistics Canada already has considerable experience and success in using aerial photographs in the National Farm Survey;
- (d) the complex process of digitizing crop boundaries can use existing technology although further technological developments are expected before 1991;
- (e) by splitting off the land-based crops module from the regular census, the timeliness and continuity of the census releases are not jeopardized;
- (f) this approach utilizes existing expertise within Statistics Canada, such as experience in handling complex data bases. The Geography Division has experience in dealing with spatial data and would play a major role in new systems research;
- (g) the Census of Agriculture already uses township plans with a quarter section grid system to ensure coverage in the Prairies. The enumerator is required to put the questionnaire number in each quarter section. It is not a big step therefore to change the system, and ask the enumerator to indicate the crops or land use instead.

### **Limitations of the Proposed System**

The move to a land-based census approach and to a different method of data collection is not without some potential weaknesses.

- (a) There will be an additional expense in using this dual approach. It is not readily apparent whether the increased cost of enumerator time completing the dual module will be substantially larger than that in completing the standard census questionnaire. It would be difficult to outline accurately the net cost at this time or to estimate the cost-recovery benefits.
- (b) There will be some differences between the crops data collected on the regular census and the crops data collected by the mapping system. These differences will have to be explained to the public although the new method could conceivably confirm the overall accuracy of the current census system.

- (c) The processing of the data, aggregation and production under the new system will be slower than the regular system. The speed of this process will be a function of resource allocation.
- (d) Since this technique is radically different from the existing system, there will be operational problems. Extensive field testing and operations testing will be essential.
- (e) The computer capacity required for a land-based data bank and analysis system will be substantial.
- (f) If there was a demand for land-based allocation of other census variables like livestock and farm income and expense data; this could be researched and a modeling approach studied.

### Research on Operational Problems

The research necessary to implement this system could be broken into the following four categories:

- (1) Data collection:      what geographic system is the best to use;  
                                     what coding and identification systems should be employed;

what are the most current developments in data analysis and digitizing equipment; field testing.

- (2) Data linkage:      what data capture system is the most efficient;

how will the data be digitized;  
which data processing procedure;  
what data verification system.

- (3) Data storage:      what type of data storage system;  
                                     what type of electronic handling;  
                                     how much computer storage is required.

- (4) Data dissemination:      what level of disaggregation should be developed;  
                                             which are the important data variables;  
                                             what type of access systems are needed by users;  
                                             what are the confidentiality problems.

# CONFIDENTIALITY OF CENSUS OF AGRICULTURE DATA

RICK BURROUGHS

AGRICULTURE AND NATURAL RESOURCES DIVISION  
STATISTICS CANADA

MARY MARCH

CENSUS AND HOUSEHOLD SURVEY METHODS DIVISION  
STATISTICS CANADA

## Introduction

There are a number of clauses in the Statistics Act that deal with the disclosure issue. The one which applies to the Census of Agriculture comes from Section 16, paragraph (1) (b). It declares that: "No person who has been sworn under section 6 shall disclose or knowingly cause to be disclosed, by any means, any information obtained under this Act in such a manner that it is possible from any such disclosure to relate the particulars obtained from any individual return to any identifiable individual person, business or organization". Applying this principle to a given statistic, one at a time, is straightforward assuming that the analyst has a detailed knowledge of the data from which it is derived and the time to consider it all. However, a good deal more effort is required to apply this principle to a large multivariate data base with a near infinite number of possible retrievals to consider.

A fully satisfactory treatment of this data base should:

- (a) ensure that the Statistics Act confidentiality requirements are met by all data releases. In other words, no press release, printed publication, special tabulation or data file released to users outside of Statistics Canada should disclose the particulars of any individual agricultural operation or operator (or result in a disclosure when combined with earlier releases);
- (b) maximize the amount and the quality of information made available. In other words, ambiguities introduced in order to prevent disclosure should be minimized in order to avoid destroying the usefulness of the estimates provided; and

- (c) be implemented at a reasonable cost.

## Nature of the Data Base

In order to appreciate the confidentiality problem posed by the Census of Agriculture, a basic understanding of the nature of the data base is essential. The 1981 Census of Agriculture has been chosen as the example since it is the most recent and therefore the most likely of available data bases to resemble the 1991 version. Since the number of farms is declining and likely to continue to decline, the problems inherent in the 1981 Census will probably be more serious by 1991.

The 1981 Census of Agriculture data base is composed of about 318,000 records (census farms). Each record may contain as many as 278 fields of data or variables. However, not all variables apply to each census farm; for example, a variable like the number of chickens will have a value only for those records (census farms) where chickens are present. Conversely, a variable like the age of operator requires that a value exist on all records.

Thus, the records generally contain fewer than 278 fields. The average record contains 14 fields of numeric and geographic identifiers, 18 fields of qualitative<sup>1</sup> information and 42 fields of quantitative<sup>2</sup> information, for a total of 74 fields.

The standard outputs include a geographic breakdown of all variables to the province, census division and census consolidated subdivision, and seven cross-classifications of the more important variables at the province level. All other output is retrieved by user-specified request. There have been over 1,000 such requests since the release of the 1981 base in July of 1982.

<sup>1</sup> The term qualitative is used here to describe a variable which is composed of codes classifying the farm according to some attribute or quality, for example, farm type.

<sup>2</sup> The term quantitative is used here to describe a variable which is made up of individual totals or quantities relating to the farm, for example, number of pigs.

## Rule of Ten Farms

The procedure most frequently used in recent censuses is called the Rule of Ten Farms. It states that data may not be released for a geographic area of less than ten farms, but will be combined with the data from geographically adjacent areas until the total number of farms is ten or more. Table 1 demonstrates the application of this rule to a set of fictitious data.

The impact of this rule can be illustrated by measuring its effect on the standard geographic levels used in the census. Four census divisions of the 257 in Canada with farms have less than ten farms; 307 Census Consolidated Subdivisions (CCSs) of 2,475 with farms have less than ten; and over half of the 13,000 enumeration areas with farms have less than ten.

This rule is relatively inexpensive to implement and is generally, although not always, well accepted by the user community. Although it tends to cover a lot of possible disclosures, it may also ignore some obvious disclosures and at the same time needlessly prevent the release of acceptable statistics as well. In Table 1, the number of pigs in CCS 3 is sensitive and hidden,

the number of pigs in CCS 2 is also sensitive but ignored, while the number of pigs in CCS 1 is not sensitive but has been hidden unnecessarily.

## Suppression of Counts

The Suppression of Counts technique is not currently in use although it has been used as recently as 1976. Simply put, a decision is taken not to release the count of the farms reporting a given variable below a given geographic level. This was generally done for quantitative variables that are relatively infrequent such as certain horticultural crops. Table 2 illustrates the technique using the same fictitious data as Table 1. The decision was taken to not release the farms reporting pigs at the CCS level because of the frequency of disclosures at that level. In this case, the data for CCS 2 and CCS 3 might be considered to be disclosures while the data for CCS 1 are not.

This rule again is relatively inexpensive to implement and removes the data that tend to identify a disclosure. The counts are not disclosures in most cases, so the cost in data loss terms is relatively high. The big problem with this rule is that it doesn't remove the disclosures, it just makes them more difficult to find.

**TABLE 1. Rule of "Ten Farms"**

**BEFORE RULE OF TEN FARMS**

	Number of Farms	Total Sales \$000	Number of Pigs	Farms Reporting Pigs
CCS 1	6	27	34	4
CCS 2	15	105	1,250	2
CCS 3	7	22	4	1
Census Division	28	154	1,288	7

**AFTER RULE OF TEN FARMS**

	Number of Farms	Total Sales \$000	Number of Pigs	Farms Reporting Pigs
CCS 1 <sup>1</sup>	-	-	-	-
CCS 2	15	105	1,250	2
CCS 3 <sup>1</sup>	13	49	38	5
Census Division	28	154	1,288	7

<sup>1</sup> Six farms from CCS 1 have been combined with CCS 3.

**TABLE 2. Rule of "Suppression of Counts"****BEFORE SUPPRESSION OF COUNTS**

	Number of Farms	Total Sales \$000	Number of Pigs	Farms Reporting Pigs
CCS 1	6	27	34	4
CCS 2	15	105	1,250	2
CCS 3	7	22	4	1
Census Division	28	154	1,288	7

**AFTER SUPPRESSION OF COUNTS**

	Number of Farms	Total Sales \$000	Number of Pigs	Farms Reporting Pigs
CCS 1	6	27	34	-
CCS 2	15	105	1,250	-
CCS 3	7	22	4	-
Census Division	28	154	1,288	7

**Rule of 3 for Distributions**

The Rule of 3 for Distributions is another rule that is not currently in use but was around in 1976. This rule is applied to frequency distributions of quantitative variables. The general enunciation is as follows: if the count in the highest class containing a value is less than 3 and the count in the next lower class is 0, then combine the two highest classes containing values and change the range for this class to read as equal to or higher than the minimum of the collapsed classes. The same collapsing will apply to an associated total. Table 3 illustrates this procedure using a frequency distribution of total sales taken from the same census division data in Tables 1 and 2.

Again, this is a relatively inexpensive procedure to implement. For the upper classes of frequency distributions it is probably quite effective, especially when used with a carefully selected set of class limits. It does not cover problems in the middle or lower classes where it is argued that disclosures are not identifiable or even trivial.

**Suppression of Totals**

This type of disclosure prevention has been in use in the Census of Agriculture in a variety of forms for several past censuses, although 1981 was the first time it was used in a systematic way.

The basic procedure involves suppressing totals of quantitative variables whenever the total is

dominated by a small number of farms. Suppression of a total will usually involve the suppression of one or more complementary totals to avoid a residual disclosure. The concentration rules used to define a sensitive total vary significantly. The two rules used in the 1981 Census of Agriculture were the so-called "Duffett Rule" and the "Less than 3 Units Reporting Rule"

Table 4 returns to the fictitious census division of earlier tables to illustrate how this technique would affect the data.

If one studies Table 3 and Table 4 together, one becomes suspicious that there is one big pig farm with sales of \$87,000 located in CCS 2. Assume this to be true. This creates sensitive totals for pigs and sales in CCS 2 and for pigs at the census division level as well. A complement will therefore have to be chosen from among other census divisions. The number of pigs in CCS 3 is a sensitive total obviously, while pigs in CCS 1 and sales in CCS 3 are complements.

This type of procedure is generally considered to be quite effective. The principal problem stems from the large quantities of resources, both computer and personnel, required to implement the procedure. The development of the CONFID software in the department made a systematic and comprehensive approach to analysing the confidentiality problems of the data base possible for the first time in 1981.

**TABLE 3. Rule of "3 for Distributions"****BEFORE RULE OF 3 FOR DISTRIBUTIONS****CENSUS DIVISION**

<b>Farms Reporting Total Sales of:</b>	<b>Number of Farms</b>	<b>Total Sales \$000</b>
Under \$1,000	11	8
\$1,000 to \$4,999	12	35
\$5,000 to \$9,999	4	24
\$10,000 to \$24,999	0	0
\$25,000 and Over	1	87
<b>Total</b>	<b>28</b>	<b>154</b>

<b>AFTER RULE OF 3 FOR DISTRIBUTIONS</b>		
<b>Farms Reporting Total Sales of:</b>	<b>Number of Farms</b>	<b>Total Sales \$000</b>
Under \$1,000	11	8
\$1,000 to \$4,999	12	35
\$5,000 and Over	5	111
<b>Total</b>	<b>28</b>	<b>154</b>

**TABLE 4. Rule of "Suppression of Totals"****BEFORE SUPPRESSION OF TOTALS**

	<b>Number of Farms</b>	<b>Total Sales \$000</b>	<b>Number of Pigs</b>	<b>Farms Reporting Pigs</b>
CCS 1	6	27	34	4
CCS 2	15	105	1,250	2
CCS 3	7	22	4	1
Census Division	28	154	1,288	7

<b>AFTER SUPPRESSION OF TOTALS</b>				
	<b>Number of Farms</b>	<b>Total Sales \$000</b>	<b>Number of Pigs</b>	<b>Farms Reporting Pigs</b>
CCS 1	6	27	-	4
CCS 2	15	-	-	2
CCS 3	7	-	-	1
Census Division	28	154	-	7

However, the expense limited the analysis to 64 variables of the 246 quantitative variables and was used only on the standard outputs. The user-specified outputs apply the "Less than 3 Units Reporting Rule", by hand, to these same 64 variables.

There is another price to pay in terms of data loss. The number of complements in most retrievals exceeds the number of sensitive cells and increases in geometric proportion with the number of dimensions in the table. This is especially disconcerting when a trivial but sensitive total sets off a chain of complements which exceeds the total of the sensitive cell by many times.

### Random Rounding to Base 5

Random Rounding to Base 5 is a technique that has been used in the 1971 and 1981 Agriculture Population Linkage data bases.

The technique used on these bases involved rounding the count of farms in a retrieval either up or down to the nearest multiple of 5, randomly and with a known probability. Then any data associated with those farms in the retrieval are adjusted in the same proportion to maintain the average. Table 5 illustrates what this procedure would do to the census division data used previously.

Random rounding is implemented at virtually no cost as long as the version currently available in the STATPAK-TARELA software is used. Its other advantage is that the resulting data at least appear to be devoid of disclosures. These advantages are offset by the fact that, regardless of appearance, random rounding does very little about dominance problems.

This should be evident from Table 5. The big pig farm is not as evident in the data after rounding but its existence can still be inferred from the census division total. The other problem which can be seen in the table is the distortion of the raw data. This distortion is significant only when the farm counts are low, but this situation is quite frequent with agricultural data, and many uses of the data would have to be foregone if this procedure became general practice.

### The 1986 Census of Agriculture Confidentiality Procedures

In 1986, the main objective will be to do the best possible job of disclosure prevention given current cost constraints. Since system resources are in short supply, disclosure prevention procedures will only use systems such as CONFID or random rounding that are already available. The only systems to be developed will be those necessary to enable a reasonably efficient interface between available confidentiality software and Census of Agriculture systems.

**TABLE 5. Rule of "Random Rounding"**

#### BEFORE RANDOM ROUNDING

	Number of Farms	Total Sales \$000	Number of Pigs	Farms Reporting Pigs
CCS 1	6	27	34	4
CCS 2	15	105	1,250	2
CCS 3	7	22	4	1
Census Division	28	154	1,288	7

#### AFTER RANDOM ROUNDING

	Number of Farms	Total Sales \$000	Number of Pigs	Farms Reporting Pigs
CCS 1	10	45	43	5
CCS 2	15	105	0	0
CCS 3	5	16	20	5
Census Division	25	138	1,840	10

Confidentiality procedures will be applied to all products of the 1986 Census including user-specified tables and micro-data files. These procedures will include measures designed to avoid residual disclosures when user-specified tabulations are combined with previously available data. We intend to minimize chances of a disclosure by trying to use available procedures as judiciously as possible.

#### Alternatives for 1991

Neither of the confidentiality procedures used in the past nor even the approach to be used in 1986 are fully satisfactory. Even CONFID, which quite effectively avoids publication of sensitive information, sometimes removes more information than seems necessary (it is hard to accept the suppression of non-sensitive estimates for important commodities in order to avoid disclosure of data for the one farmer who produces small amounts of a commodity which is rare only locally). Implementation of CONFID also proved to be a large task in 1981 which had to be fitted into a tight production schedule. None of the procedures other than random rounding is at all effective in avoiding residual disclosures in user-specified requests.

The disclosure prevention procedure is an important part of the survey process. Therefore, effort should be made to ensure that the procedure is effective and of comparable quality to other parts of the survey operation.

As part of 1991 planning, alternatives to present procedures should be sought. These might be no more than modified versions of procedures used in the past. For example, random rounding is suitable for some variables with well-behaved distributions, and it could be quite effective if it were controlled in some manner to avoid excessive distortions while at the same time introducing sufficient ambiguity. The CONFID procedure is also attractive because of its effectiveness in preventing disclosures, particularly those involving commodities where dominators are common. Perhaps it need only be used selectively. Also worthy of consideration is an entirely new approach. For example, some consideration might be given to a procedure that would be applied at a micro level to all or perhaps only some records. The confidentiality problem is, after all, a micro level one in that individual particulars must be protected. Why not have some method of distorting or collapsing the micro information?

Whatever approach is used for 1991, however, it should be one that we are prepared to use in future censuses. In other words, we would like to

end the continuing debate and changes in Census of Agriculture confidentiality procedures from census to census.

#### Disclosure Prevention Procedures for Future Censuses of Agriculture

Confidentiality procedures are now a major issue within Statistics Canada. In fact, a special task force has been named and assigned the responsibility of studying procedures currently in use and making recommendations for the future (with respect to standardization, mechanisms to be developed, etc.). Possibly, as a result of the task force's work and of other initiatives, guidelines will be defined that can be applied directly to the Census of Agriculture.

If suitable and easy-to-use mechanisms for the prevention of disclosure also become available in the meantime, it may be possible to use one or more of them. This of course assumes that one of the mechanisms will be an optimum procedure appropriate for the Census of Agriculture, data which are in many ways different from other Statistics Canada products.

To be realistic, even if an "off-the-shelf" solution is used, at least some modifications to it will be necessary. Experience with the CONFID collapse and suppress software used in 1981, for example, has shown that incorporation of an already available confidentiality mechanism can be a major task, particularly if a large number of estimates are involved and if interfaces with other parts of the system are not smooth. Analysis must be undertaken to ensure the confidentiality procedures can be integrated with the data producing system. Ideally, they should be as much a part of the system as imputation is, for example.

Careful study of the Census of Agriculture situation is necessary in order to determine an appropriate procedure for confidentiality protection. The types of disclosures that will occur are dependent on the geographic and size distributions of the data and on the nature of the tables that are produced (including the number of dimensions, the number of levels within each dimension, presence of corresponding frequency counts with quantitative data, whether categories add up, etc.). The choice of procedures used to prevent disclosures also must take into account the ways in which the data are to be used.

It should also be borne in mind that all data releases from a given census are not simultaneous and cannot be predetermined. Special user-requested tabulations and data files must be handled appropriately in order to avoid residual

disclosure when they are used in combination with information that has already been made available.

It is recommended that analysis and development work begin well before 1991 if a satisfactory procedure is to be in place by that time.

### **Proposed 1991 Development Activities and Their Schedule**

If a confidentiality procedure that is effective from cost and quality points of view is to be incorporated into the 1991 Census, two activities must be completed before the date when the 1991 team begins preparation of specifications for the census systems (approximately three years before Census Day).

The first activity, analysis of the Census of Agriculture situations where disclosures are possible and of quality requirements of users, should be completed as soon as possible. Some

analysis has already been done as part of the development of systems for 1981 and 1986. Further investigations should be carried out mostly by subject-matter staff (with participation of methodologists who can help in the study of data distributions, for example). The latest possible date for completion of this task would be June 1987 (leaving only a year for resulting research activities before the 1991 specification process begins).

The second activity, incorporation of an appropriate procedure into the Census of Agriculture methodology, should be completed in the following year. Ideally, the procedure should be part of the census system in the way that imputation is. In fact, the procedure might even be part of the imputation process so that we would end up with three bases: the unimputed base, the imputed base and the base adjusted against disclosures.<sup>3</sup> It is probably more realistic, however, to expect that the procedure will be incorporated into the output program.

<sup>3</sup> This supposes that a micro-level procedure might make sense. We impute for non-response and we would also impute as a confidentiality procedure (i.e. replace actual responses or move them).

<b>Proposed Development Schedule</b>	
<b>Activity</b>	<b>Date</b>
1. Define Census of Agriculture Needs for a Confidentiality System.	June, 1987
2. Finalize Methodology.	June, 1988



## **Résumé of Discussion**

**Mr. Jones**, manager of the 1986 Census of Agriculture, reviewed the three presentations.

### **Mail-out/Mail-back Methodology**

The first presentation, on the U.S. Census of Agriculture, was of particular interest to this group because it gave some idea of the problems likely to be encountered if Canada's Census of Agriculture were to be done using a mail-out/mail-back methodology. The advantages of being associated with the Census of Population should not be overlooked. These include tremendous cost-savings as well as being considered so small relative to the much larger Census of Population that the Census of Agriculture is rarely subjected to drastic budget cuts. It would not be wise to end the relationship between the two censuses.

Similarities between the U.S. and Canadian censuses include often identical issues arising that are related to content and concepts and the two publicity programs.

### **Proposal for a Land-based Census of Agriculture**

The second presentation was a proposal for a land-based Census of Agriculture possibly making use of satellite data collected during the same time period. One weakness of such a procedure would be the risk of equipment failure.

A land-based census has been previously suggested by users from Environment Canada, Agriculture Canada, the Saskatchewan Department of Agriculture and the research community. Since one purpose of the Census is provision of small area data and since the present method of collection results in distortions at lower geographic levels, it is clear that research should be done to develop better methods.

The suggested methods of making data land-based will mean a substantial increase in costs. Also, one risk of improving the quality of some variables at a

small-area level, as suggested in the presentation, is that users will continue to demand more detail for more variables, thereby further increasing the cost of the census. Although some of the costs can be recovered from the users, there may not be sufficient funds to cover all of the increase.

As stated in the earlier presentation on extension of mail-back, there is a need to decrease rather than increase field costs of surveys. To increase the complexity of field procedures in these times appears to be a backward step.

### **Confidentiality Procedures**

Census of Agriculture confidentiality procedures were the subject of the third presentation. These procedures have been changing for every one of the last three censuses and still the definition of disclosure is not clear. There is a need to reduce both the risk of disclosure (however it is defined) and the cost of the procedures used.

In developing a new system, care should be taken to ensure a reasonable balance between protection of confidentiality of supplier data and the quality of information provided to users.

A generalized system usable by several surveys is desirable but not yet forthcoming. An attempt to develop a specialized system for the Census of Agriculture may result in a management realization of this need and funds may instead be made available for the generalized system. In any case, it is worthwhile to begin development of the system for the Census of Agriculture.

### **Summary**

In summarizing, **Mr. Jones** noted that research work in the two areas mentioned should be started as soon as possible. In addition, research related to the proposed extension of mail-back is also necessary. One possibility for making the idea more workable is the addition of a farm operator question to the population Forms 2A/2B.



## **Résumé of the Question Period**

*The Census of Agriculture session was followed by a question period where several concerns relating to the issues were raised and discussed by the participants.*

### **Need for Small Area Data**

*The need for small area data was questioned. Are they needed for their own sake or because users want to be able to re-aggregate them at will? Are only limited sets of re-aggregations desired? If so, cannot these different aggregations be provided instead of the smallest level data?*

*Very few user requests are handled by the U.S. Census of Agriculture. The smallest areas for which data are requested are usually aggregations of at least two or three counties. Special tabulations are usually very expensive (costing \$50,000 on average) and as a result there have been fewer than 100 such requests for 1982 Census data.*

*In Canada, there are many requests for special tabulations of Census of Agriculture data. Tabulations are relatively inexpensive (at an average cost of \$300). Some users prefer to aggregate data themselves while others request totals for conservation areas, drainage areas, a federal riding, etc. All requests that do not break any confidentiality rule are handled. There is a demand to be increasingly flexible.*

## **Relationship Between the Censuses of Population and Agriculture**

*The relationship between the Censuses of Population and Agriculture was discussed. Clearly, planning work would have to be done assuming that the two censuses would continue to be carried out together. A change in the nature of the relationship might, however, be unavoidable, if it became necessary for the Census of Population to move toward less enumerator contact. One possibility would be to reduce the Census of Population involvement to a coverage function while the Census of Agriculture would become a follow-up survey.*

### **Confidentiality Issue**

*The remainder of the discussion was on the topic of confidentiality. Generally, there was a feeling that too much information was being suppressed and that efficient and effective procedures were needed for the handling of special requests. Although, methodologists from different countries were beginning to share information on the subject of confidentiality procedures, they were not yet working together. It was noted that the United States Bureau of the Census and Statistics Canada were still in the process of trying to coordinate their work internally. It was suggested that the definition of a disclosure in the Census of Agriculture would likely be consistent with definitions used when releasing business data.*



## **SESSION: CENSUS AUTOMATION**

Chairperson:     Martin Podehl  
                         Informatics Services and  
                         Development Division  
                         Statistics Canada

Thursday, October 10, 1985



# AUTOMATION PLANS FOR THE 1990 U.S. CENSUS OF POPULATION AND HOUSING

PETER A. BOUNPANE

U.S. BUREAU OF THE CENSUS

## Introduction

Automation is one of the key areas we are examining as we plan the 1990 U.S. Census of Population and Housing. Automating many of the census tasks performed clerically in 1980 and previous censuses can help us to take the census more quickly, allowing us to meet our legal mandates for releasing apportionment and redistricting counts and to release other data products in a timely manner. Automation could also help us introduce cost-efficiencies into many areas, improve accuracy, and also allow for better control of the census process.

Traditionally, census data collection and much of the census data processing (e.g., questionnaire check-in against the address control list, edit of questionnaires for completeness, and coding of handwritten responses) have been paper- and people-intensive tasks. The use of automated equipment can help to deal with the mountains of paper and the thousands of clerical tasks in a much more efficient and controlled way. Hiring, training, and finding space for all the people who have been needed to perform the numerous operations in past censuses have taken a lot of time and cost a lot of money. While the 1990 census will also likely require a large number of temporary workers, we are looking at ways to cut down on the number of labor-intensive activities and to use automated systems to control the census process.

We have been working on our automation plans for some time now. We have tested some new approaches in our test censuses this year in Tampa, Florida, and in Jersey City, New Jersey, and will conduct further tests of automation next year in parts of Los Angeles County, California, and in several counties in East Central Mississippi. These tests are very important as laboratories where we can try out optional approaches, and I will be referring to them throughout this paper.

Since we are contemplating significant changes in automation for 1990, I will first describe how the 1980 census was taken so that the departures can be more easily understood.

## 1980 Census

The 1980 census was taken using the mail-out/mail-back procedure in areas of the country containing 95 per cent of the population. We purchased address lists for some of these areas and listed addresses ourselves elsewhere; in all cases, the address lists were then checked and updated by the U.S. Postal Service and our own field personnel. The USPS delivered questionnaires to each housing unit a few days before Census Day and householders were asked to fill them out and mail them back to a temporary census district office on April 1. The aim of this approach was to complete as much of the census as possible by the less costly mail method and then to do the costly and time-consuming followup of those housing units for which no questionnaire was returned. We had received questionnaires for about 83 per cent of the occupied households within two weeks of the mailout. A large work force (270,000 at peak) personally visited nonresponding housing units and vacant units. In sparsely populated parts of the country where mail-census procedures were not suitable, census enumerators went door-to-door to take the census.

We set up 409 temporary district offices to carry out data collection. For each office, a large number of clerks were hired to make changes (additions, deletions, corrections) to the address lists, check in mail-returned questionnaires and edit the questionnaires for completeness and consistency, assign housing units for followup, monitor the enumeration of the nonresponding units, and tally preliminary counts. The majority of these operations were done manually. Many of these operations can be considered "processing," but processing did not begin in earnest until the collection offices completed their work, closed, and shipped their questionnaires to one of three processing centers. The offices generally closed from 5-7 months after Census Day.

At the processing centers, the questionnaires were microfilmed and the data read into computer by FOSDIC. Though processing center operations were largely automated, written entries for many questionnaire items (e.g., ancestry and occupation) were given numerical codes manually prior to computer processing. These coding operations required a large clerical work force.

This system worked very well considering the amount of manual work involved and the sharp division between data collection and data processing. The Census Bureau met its legal deadlines for the release of apportionment and redistricting counts; many of the small-area data, such as block counts, were out earlier than for the previous census; and many more data, especially for race and Spanish-origin groups, were published. Still, we did not release some of the data products, particularly those based on the sample questions, as quickly as planned. (This delay was due in part to budget problems that forced us to cut staff and temporarily suspend sample coding operations.)

For the 1990 census, we want again to meet our legal deadlines and we want to release other data products more quickly than ever, as well as keeping costs reasonable and making the counts as accurate as possible.

### **Automation Plans for 1990**

We have identified a number of areas that are candidates for automation, and have already begun to test some of them.

One of these areas is geography. Geographic materials are essential to a successful census for two reasons: First, having correct and legible maps helps our enumerators find every housing unit so that we have a complete count; and second, having correct boundaries and geographic information helps us assign each housing unit and the people who live there to the appropriate land area. One of our problems in the 1980 census was that our geographic materials, including the maps, were produced in separate operations involving a great deal of clerical work. This process was slow and error-prone, leading to delays in production and errors and inconsistencies in some of the products.

For 1990 we are automating our geographic support system, which we are calling TIGER (Topologically Integrated Geographic Encoding and Referencing system). TIGER will integrate into one file all the geographic information that was produced in separate operations in 1980. This will allow us to produce the geographic products and services for 1990 from one consistent data base, and will help us avoid some of the 1980 census delays and inaccuracies. Having the computer generate maps that match the geographic areas in our tabulations will be a big improvement over the clerical operations of the 1980 and earlier censuses. Another paper at this conference will describe the automated geography system in more detail.

Another improvement planned for the 1990 census is the development of an automated address control file. In 1980, although the initial control list of addresses was computerized, changes to the address file during the census were made manually. For 1990, we will have continuous access to the automated address control file so that we can keep the list current. We have already implemented an automated address control file successfully in our 1985 test censuses and will conduct further testing.

With an automated address file, it will be much easier to determine whether or not we included a specific address in the file. It also will be possible to update the file where we missed an address in earlier operations. We can use bar-code technology for computer check-in of the questionnaires. As a result, it will be easier for our enumeration staff to identify the addresses for which questionnaires have not been returned, and we could send reminder notices to those addresses, thus reducing further the number of nonresponding housing units where we need to send enumerators. Finally, with an automated address list, we can update the list and use it in future Census Bureau operations. In our 1985 test censuses, we successfully implemented an automated address control file, automated check-in, and the use of reminder cards.

One of the most promising ways to take advantage of automation in the census, and our biggest challenge, is to convert the data on the questionnaires into a computer-readable format earlier in the census process than in past censuses. This approach is essential if we are going to take full advantage of automation and release data products quicker. For 1980, the data conversion did not begin until after the temporary census offices closed and shipped their questionnaires to one of three automated processing centers. For the 1990 census, we want to begin converting data simultaneously with the collection phase. This early start (5-7 months ahead of the 1980 schedule) will allow more time for review and correction and will enable the computer to assist in certain census operations. It will contribute to tighter control of field followup assignments and allow early identification of enumeration problems. Also, computer records of questionnaires could serve as backups to the originals in case they are accidentally destroyed.

Although there is agreement that we should implement earlier automated processing for the 1990 census, there are two major questions we still must answer. Where will the automated processing be conducted, and what technology will be used to convert the questionnaire data into computer-readable form?

With regard to the first question, it is helpful to consider two broad scenarios for accomplishing this early data conversion. Under one scenario, there would be combined district and processing offices, which would carry out both automated processing activities and field followup. It is very unlikely we would use "combined" offices for the entire country because of difficulties building, installing, integrating, and monitoring 500 separate data processing systems; however, they may be used in more rural areas. We will be testing a "combined" office in our 1986 test census in Mississippi.

Under the other scenario, we would have separate processing and district offices. Here, the processing offices would receive the mail-returned questionnaires from the public, check them in automatically, convert the data to machine-readable format, and perform automated editing of the questionnaires. The district offices would be responsible only for contacting households to follow up missing or incomplete questionnaires. We tested this plan in our 1985 test censuses, with collection offices in Jersey City, New Jersey, and Tampa, Florida, and processing in our permanent processing office in Jeffersonville, Indiana, and it worked quite well. In our 1986 test census in Los Angeles County, we will again use separate district and processing offices, but the processing office will be within the same metropolitan area as the district office. It is unlikely we would use "separate" offices for the entire country because of the communications and logistics problems that would arise if the processing office were a long distance away from the district office.

Having combined processing/district offices in parts of the country with low population density and separate processing and district offices in the more urban areas is a likely option for the 1990 census.

In addition to deciding where to convert the data to computer-readable format for 1990, we must also determine how to do so. In 1980, after we completed collection activities, we entered the data from the questionnaires onto computer tapes by microfilming the questionnaires (after clerically coding write-in responses), and then reading the microfilm with an optical scanning device (FOSDIC - Film Optical Sensing Device for Input to Computers). The choices for 1990 are basically among three technologies or various combinations thereof. We can continue to use the film-to-tape process as in 1980, but with newer and better equipment. We can try to eliminate the microfilming step and read the questionnaires directly as college aptitude tests are processed using optical mark recognition technology. Or we

can enter the data by keying. Keying for all data conversion in all processing locations is unlikely, but we will need to use it extensively for entering into the computer address information and the written answers on the questionnaires.

In our 1985 test censuses, we used the Optical Mark Reader and keying approaches. In the 1986 tests, we will use keying and the film-to-tape method. Although there were some problems, the Optical Mark Reader worked well enough for us to consider the possibility of testing it further. One problem is that the OMR reader available at the time could process only 8-1/2" x 11" paper. This meant that we had to squeeze the entire short form questionnaire onto one small page. This created design restrictions that may have contributed to response problems on several questions. If a reader that can process 11" x 17" pages can be developed, we will do a "hot house" test next year.

The issue of data conversion methodologies is related to, but not dependent on, the office structures discussed above. A decision on equipment also involves many other considerations such as the content and appearance of the questionnaires and the ease with which people can complete them; the reliability and availability of the equipment; the staffing requirements imposed by the equipment both in terms of numbers of people needed and the technical sophistication those people must have; and the cost and maintainability of the equipment.

Another paper at this conference will discuss the issue of concurrent (or decentralized) processing in more detail. We will increase or improve automation in other areas to help speed up the census and make it more accurate, and I will discuss briefly a few of these areas.

One area is questionnaire edit. Edit is a repetitive and monotonous job better suited to computers than people. Entering data from the questionnaires into the computer earlier in the census process will allow computer editing of the questionnaire data earlier than ever before. These edits will check the completeness and consistency of the data. In 1980, the questionnaires were manually edited in the district offices, basically to check that they had been answered completely; then, once the questionnaires went through the FOSDIC machines, the computer edited them for completeness and consistency. For 1990, most of the manual editing would be eliminated, resulting in speedier, more consistent, and more accurate editing. Decentralizing the computer edit operation would also allow us to recontact the respondent, if necessary, which was not possible in the 1980 census.

Another promising automation technique relates to the coding of handwritten entries on the questionnaire. In 1980, manually coding the handwritten entries on questionnaires involved a large, time-consuming, and costly clerical operation. For 1990, we will key handwritten responses into the computer and specially developed software will assign the appropriate computer-readable codes. We cannot eliminate all clerical involvement in coding, because some handwritten responses will be incomplete or uncodable and will have to be handled by our referral units. We will, however, be able to significantly reduce the amount of manual work and, thus, save time and money and improve the quality of the data. We are planning to test some aspects of automated coding in our 1986 test censuses.

We will also use automation to help us plan and monitor the census. The Census Bureau is developing an elaborate automated management information data base to see that we meet key dates in making decisions about the shape of the 1990 census. The management information system was in place to help us keep track of operations for our 1985 test censuses and in helping us plan our 1986 test censuses. In addition to serving as an aid in planning the 1990 census, the management information system will give us up-to-the-minute cost and progress data so that we can monitor actual 1990 census operations. In 1980, cost and progress reports were not integrated with other management reports, and some of the cost and progress information was several days old by the time managers received it.

Automation will help us control and monitor many other administrative functions. We will have an automated payroll system, as in 1980. And for 1990, we will also have, on a micro-computer, a new automated employee file that will help us organize needed information about our large temporary work force. (We did this in our 1985 test census.) For instance, we will know whether we are meeting our hiring goals in each enumeration area and we can use the file to help us make enumerator assignments. We will also have a new automated inventory control system to manage the procurement and distribution of the large volume of specialized supplies needed to take the census.

Finally, we are looking at automation of our tabulation and publication operations for the 1990 census. Our tabulation system was fully computerized for the 1980 census, but for 1990 we expect to take advantage of advances in data base software to make improvements in the system. We

also want to use the computer in our analytical review of the tabulated data, which was conducted manually in 1980. This review, which looks for errors and anomalies in the data, is essential to maintaining the quality of our data products. Using the computer will speed up and improve this analysis.

New automation techniques will also play a part in the dissemination of our data products for the 1990 census. While the Census Bureau will continue to produce paper reports and large summary computer tape files, we must also address the needs of small computer users who will want products on floppy disks. Another new development we will consider for 1990 will be an online data base in which users can access summary data from their office computers using a telephone hookup. The Census Bureau has already implemented such a system, called CENDATA, on a limited basis. There may be other developments in the next few years - such as improvements in laser disks - that we will be able to take advantage of for the 1990 census. Fortunately, our final decisions on tabulations and data products can be made later in the decade, so we can take advantage of new technologies.

There is a sense of excitement at the Census Bureau about these automation possibilities, but some words of caution should be added. Whatever systems are developed must be simple, because they will be operated by a temporary work force with minimal training. The systems must be fully tested, proven to be reliable, and essentially "fail safe" to avoid crippling breakdowns. The equipment must not be unreasonably expensive and should either continue to have value to the Census Bureau or be marketable to someone else upon completion of the census.

Most of all, as we look to increasing automation in the census, we must take care to ensure that the confidentiality of the data we collect is maintained both in fact and in appearance. Only by maintaining the confidentiality of the census process can we ensure a high level of public trust and cooperation. The Census Bureau is proud of its record of protecting confidentiality and is constantly looking for ways to maintain and improve that protection. The Census Bureau does not release data about individuals to anyone, including other Federal government agencies. But the sometimes menacing implications of technology require that we increase our efforts to convince individuals that they cannot be harmed by answering the census and that the information they provide is strictly confidential by law.

## Closing

We must make decisions on these two major questions (where and how) related to data conversion by September 1986, so that we can begin the process of procuring equipment. Some have suggested that we should make these decisions earlier, and we will accelerate the process wherever we can. Still, we believe it is important to learn as much as possible from our test census experiences before making such major decisions. We are holding a Decennial Census Decision Conference in mid-October, 1985, to address several issues, including the "where" question.

While there are many decisions yet to be made and problems to be worked out, we have progressed far enough in our automation planning to say this: there will be significantly more automation in the 1990 census than in any previous census. We will make innovative use of automation techniques to perform data-entry earlier than ever before. We will have an automated geographic support system. We will edit questionnaires by computer. And we have already implemented an automated address control file, automated questionnaire check-in, and an automated management information system in our 1985 test censuses, and plan to have these features in 1990.



# DATA CAPTURE ALTERNATIVES

DAVE A. CROOT

CLIENT SERVICES DIVISION  
STATISTICS CANADA

## Introduction

Since the advent of the computer age, processing power has shown dramatic increases. The corresponding pressure for efficient and convenient methods of data capture to feed these voracious machines has led to improvements in methods, also at a rapid pace, if not with quite the same dramatic impact.

Data capture methods can be viewed as belonging to one of two categories; those using a human operator, and those based (almost) entirely on machines. Originally, all "business" data capture was performed by operators using machines which comprised a keyboard and a punching mechanism. The output medium was usually 80-column cards or punched paper tape. Although over time the nature of the machines used has improved, with consequent improvement in productivity, the presence of a human operator is still the main characteristic of this class of data capture.

Some early scientific uses of computers included facilities for direct data capture by some form of sensor linked to laboratory or industrial process instrumentation measuring such phenomena as temperature or flow rates. These have also improved through time, and the sensing of data by machine has been extended to the business world, that is to say, to numbers and words inscribed on paper.

Data capture, by people or machine, consists of three steps:

1. Recognition
2. Interpretation
3. Recording

## Human Data Capture

In the case of people-oriented methods, the early reliance upon holes punched in paper as the machine-readable medium has given way to the recording of data on faster and more convenient magnetic media. Improvements in communication systems enable direct transmission to distant sites, obviating the need for any intermediate storage medium where data are collected regionally for central processing. However, most

technological developments have addressed the first step in the process. With the advent of cheaper and more easily available computing power it has been possible to combine data capture with other related operations such as editing or coding. By precluding repeated handling of the same piece of paper, direct labour saving is achieved, but perhaps more significantly, gains in accuracy are also possible. One particularly impressive example of this line of development is CATI (Computer-assisted Telephone Interviewing), where respondent contact, data collection and capture, together with some editing, are all performed in a single step. The above referred improvements in the computing power purchasable at unit cost have been accompanied by similar gains in the computing power available in a given volume and weight. A processor with considerable power, ample random access storage (256K) and reasonable programming facilities can now be obtained in a unit the size and weight of a portable telephone. This availability of computing power gives data capture terminals, weighing one pound or less, the ability to capture, edit, and store enough data for several hundred survey records. Thus, it would be possible to have a mobile data capture operation proceeding for several days before down-loading the data to a larger computer became necessary.

In general, one can say that most advances in human-based data capture have taken as a premise the existence of a human-paced activity and attempted to make that activity as complete and comprehensive as possible without loss of pace.

## Automatic Data Capture

The first business example of an automatic method of data capture was the use of Magnetic Ink Character Recognition (MICR) equipment by the major banks. The initial application was high-speed reading and sorting of cheques in large clearing house operations. The same characters can still be seen at the bottom of cheques although magnetic ink is no longer used. Cheques are now scanned optically, rather than magnetically. The next development was the highly stylised optical fonts now on cheques which, as stated, closely resemble the magnetic ink characters that they succeeded. Again the pioneering users were the banks, or organisations with similar needs, to

process very large numbers of comparatively simple transactions. These earlier developments were characterised by limited specialized character sets with a high degree of discrimination between characters to permit unambiguous recognition by the scanning mechanism. With the limited resolution available, ease of machine interpretation was the paramount objective, usually at the expense of user friendliness. In almost all cases the characters were first imprinted by another machine, thus retaining control of the principal factor affecting the performance of the character scanning to follow. There has been a steady improvement in the speed and resolution of scanning and this has permitted an improving humanisation of the character sets able to be processed by automatic scanning. Scan interpretation is based upon either:

1. bit-mapping a digital image of the object character into a matrix for comparison with prescribed character matrices; or
2. curve-following the character outline for such comparison.

Clearly, improvement in resolution, whichever of the techniques is used, has a beneficial effect on the accuracy of interpretation. The early magnetic ink characters and their optical cousins were highly stylised so as to facilitate interpretation, even with limited resolution. Further advances make it possible now to scan with a high degree of accuracy type-written material with multiple styles and sizes of font (e.g., as it might arrive from a number of different respondents).

Another major development, starting in the 1950s, was the scanning of handwritten data, initially of marks only, rather than alphanumeric characters. One such development was FOSDIC (Film Optical Sensing Device for Input to Computers) developed at the U.S. Bureau of the Census, and used in Canada for the 1971 and 1976 Censuses of Housing and Population. In FOSDIC, questionnaires are first filmed. As a result, the scanning operation is comparatively positive because discrimination is of marks from transmitted rather than reflected light. Discrimination is aided by the necessity to determine only the presence or absence of marks in particular positions. Meaning is ascribed by the positions in which the marks are placed. Many questions in social surveys such as the census have a small set of possible responses and therefore lend themselves quite naturally to a multiple-choice format, with the consequent ability to represent the response by optical marking. Since it is rarely

possible, except with highly contrived question construction, to handle all questions in this manner, the capture of characters or numerals becomes necessary. Machines are now available which will read limited hand-formed characters. The vagaries of hand printing, let alone handwriting, mean that the accuracy is still suspect and usually the system requires some human supplement. When a character is scanned it is digitised, and the resultant matrix or digital representation of the image is matched by the system against a series of entries in its file of potential characters. The more powerful systems have larger files of potential characters, and can be responsive to alternate ways of representing the same character. This may be used also to give the ability to "learn" a particular alternate by adding new entries to the files based upon the material being scanned. The human supplement takes the form of giving a judgement when the system cannot match the character being scanned to any in its files. The digital representation is retained and presented for human interpretation. This is done by displaying the context, that is the recognised characters in each record containing one or more unrecognisable characters, together with the successive digital images of any unrecognisable character. The human operator, with greater mental agility than a machine, can then assign a meaning to the image. These assignments can be used heuristically, when the variants of characters are likely to prove repetitive, or simply as a step in the operational process.

Another line of development in automatic data capture has been automatic voice recognition. The preceding description of the evolution of optical character scanning could be used with little modification for voice recognition. The evolution, after a much later start, has been rapid in spite of the fact that the process is intrinsically more difficult. This is so because of the extra dimension of variability caused by the differences between male and female voices, pitch and vowel sound differences in various national and regional accents, and the fact that normal speech is even more likely than writing to be made up of a continuous stream rather than clearly separated characters. Presently, vocabularies of 500-1,000 words are processable with reasonable accuracy. Since the nature of this method lends itself to real time use, a particular system may give one the opportunity to re-articulate initially unrecognised words, more slowly. Use tends to be gravitating towards specialised applications involving immediate access to computer files as an adjunct to a relatively slow-paced activity, e.g., symptom-reporting in medical diagnosis or telephone directory assistance.

## The Choice for 1991

The essential choice for capture of 1991 Census data is between keying and a machine-based system, with the latter almost certainly involving some form of optical scanning, of either characters or positional marks, or a combination thereof. Although impressive progress has been made in this field, we remain confined to predictable separable characters rather than true handwriting. In an operation like the census it would be essential to be able to capture an extremely high percentage of data without the need for subsequent human intervention. This would require the development of powerful pattern-fitting algorithms and copious computing power. The industry appears to be on the threshold of having suitable algorithms. However, the computing power required together with high quality optics, and sophisticated mechanical paper handling equipment, are likely to necessitate a considerable capital investment in the machinery involved.

The greater the degree of automation introduced the greater tends to be the cost of the machines used; this is a particularly critical factor for an operation such as the census which is characterised by very large volume, tight deadlines and lack of frequency.

It is clearly possible to start up and shut down at least the labour part of a labour-intensive operation for a short duration activity such as census data capture. Indeed, most other census operations leading up to the automatic editing step are handled exactly that way. Furthermore, the labour cost is essentially constant if project deadlines are compressed. With a capital intensive method, deadline advancement would probably entail the use of additional equipment and almost certainly increase process cost.

We have been particularly fortunate, for the 1981 and 1986 Censuses of Population, to have a complete and highly efficient key-entry system, which was mounted from the infrastructure maintained by Revenue Canada (Taxation) for their annual processing cycle. A comparison prior to the 1981 Census projected a slight saving for keying, rather than using the FOSDIC as had

been in use for the preceding two censuses. A further saving was obtained by contracting the keying operation with Revenue Canada (Taxation) who had most of the equipment and, as stated above, all of the infrastructure necessary. In the event, the cost of keying, verification and transmission to Ottawa from regional centres was less than 50 cents per household, and the elapsed time was two to three months less than with the 1976 (FOSDIC) system.

There are interesting developments in optical scanning towards lower cost, lower performing systems in addition to the large up-market scanners. These are of a scale suitable for a work station with an operator assigning unrecognisable characters. This type of machine would support a continued regional approach to data capture, and may prove a reasonable compromise in degree of automation in the future. Another interesting development is optical disk storage. FOSDIC, as stated, requires prior filming of the questionnaires, but since there is a requirement for the census to retain a permanent image of the questionnaire, that step serves a dual purpose. With the development of high-performance scanners, the digital image of each questionnaire could be similarly retained, as the permanent questionnaire image, in addition to its role in data capture. Several commercial systems, integrating scanners with the massive unalterable storage capacity available with optical disks, together with retrieval to display terminals or printers, are appearing on the market. This line of development together with the above-mentioned lower cost scanner/interpreters might enable a mix of automatic scanning and high productivity key entry.

## Conclusion

At present costs, automatic scanning equipment could not compete with keying if the equipment was used only by the census. Progress is continuing, and there is a high probability of technical feasibility before 1991. However, it is not likely to prove economically feasible to use a capital intensive automotive method of data capture for the census unless the same equipment is also used for a large number of other surveys in order that the capital cost can be spread over a number of ongoing operations.



# DECENTRALIZING 1990 CENSUS DATA CAPTURE

ARNOLD A. JACKSON

DECENNIAL OPERATIONS DIVISION  
U.S. BUREAU OF THE CENSUS

## Introduction

The Bureau of the Census faces the 1990 Census of Population and Housing with an ambitious goal of holding the cost of conducting the next decennial census to a housing unit cost equal to, or less than, 1980 (in constant dollars). At the same time, demands on the Census Bureau for more, earlier, and higher quality data from the decennial census continue to swell. Census Bureau data users seem unfazed by shrinking real Federal resources and expect more for less as they exert pressures on all fronts.

Not surprisingly, with these competing conditions of lowering costs while producing as much or more than before, the Census Bureau has turned to automation as a potential solution. While 1990 will not represent a completely automated census, each major phase of the census-taking process is undergoing a thorough examination followed by extensive study and field testing to confirm suitable automation opportunities for 1990.

These phases are:

- receipt and check-in of mail return respondent and census enumerator filled questionnaire forms (over 100,000,000 expected in 1990);
- verification of the housing unit addresses of the returned forms by structured block level geographic segments (address control);
- editing of content and coverage;
- conversion of acceptable content to machine-readable form; and
- tabulations for publication of census data products.

In 1990, these phases would require over 60,000 clerks at a cost of close to \$1 billion without automation. These operations were carried out at three sites in 1980 to provide preliminary counts for Census Bureau Headquarters. The 7-month process was a sequential one with conversion of census data for computer reading occurring only after all forms for a district office had passed through the earlier stages in a batch mode.

From this experience which was costly, error prone, and tough to manage, the Census Bureau began to seriously look at the potential benefits that would arise from concurrent rather than sequential processing. Those expected benefits included:

- Elimination of original paper from the collection and processing offices as each form is cleared.
- Identification of questionnaires with content and/or coverage problems during the initial phases of collection to accelerate followup work and allow searches, matches, and other coverage refinements to occur in a timely fashion. These later followup efforts must commence while suitable field staff are still in the district offices.
- Immediate containment and resolution of collection and processing problems specific to certain geographic areas in time to design and launch special corrective actions.
- Completion of all field collection, processing, followups, and local reviews in time to carry out final tabulations (including any required adjustment operations) well before census counts are required for apportionment of the U.S. House of Representatives (December 31, 1990).

Accordingly, the options that allow the Census Bureau to pursue this approach to collection and processing are thought to be those termed as decentralized. Decentralization, as used in the balance of this paper, means carrying out collection and processing up to the final tabulation stages in a large (20 or more) or small (5 to 20) number of sites remote to Census Bureau Headquarters.

## The Basic Alternatives

A myriad of decentralization options potentially enhance the Census Bureau's ability to do concurrent collection and processing. Currently, a number of alternatives are being studied. The alternatives all share a certain level of costs that the Census Bureau must bear to finance the

hardware, software, space, and administrative support required for concurrent collection and processing. Thus, the real question to be answered is, "do the benefits of concurrent collection and processing justify the cost of decentralizing to implement it." We believe strongly that attempting concurrence of these operations in a centralized fashion is prohibitive in terms of their logistics and scope of the type of facility that would be needed.

This paper deals with two of the many options being considered. This paper intends to provoke a discussion today of the requirements, pros and cons, and the strategic implications of the choices we face. The two options discussed here are:

1. Full Decentralization  
(Exhibit A)
2. Regional Decentralization  
(Exhibit B)

These options were chosen for this paper for two reasons. First, the full decentralization option represents the extreme among the options being considered. It would consist of 350 to 450 offices while the other options are based on considerably fewer sites. Second, both are based upon our 1980 nationwide network of district collection offices and field regional offices. In 1980, the district offices sorted, checked in, edited, and otherwise prepared forms for data capture which occurred in one of only three sites. The permanent field regional office structure was appended with census centers to manage the district offices, each of which served roughly 250,000 households. Also, the potential benefits and the most imposing costs are clearer in these two straightforward extremes than in some of the hybrid alternatives. Following is a detailed description of the two options. Then, the pros and cons are presented for each option.

The final section of the paper deals with some strategic implications that link our analysis of the individual options for collection and processing to our 1990 census goals of timing, quality, cost, public cooperation, and avoidance of risks.

### Full Decentralization

In this option, collection activities and processing operations are combined under one roof at the district office level. These offices would carry out questionnaire check-in, address file updating, questionnaire editing, data conversion, searching and matching, and the followup operations necessary to ensure data quality including personal visits for reconciliation of sample form edit failures.

Since each district office is virtually a self-contained census center, no paper is shipped from one office to another. Data conversion would be done by direct terminal entry or desktop scanners. Logistics are fairly simple and management accountability can be focused clearly.

Field support should be at the highest level in this option since field followup assignments would be generated as the online address control file produces daily status reports by block. No time to transmit information is needed in this option. Also, Census Bureau management control of operations could be exerted through the existing regional structure. At the same time, however, the Census Bureau would be undertaking the challenging task of managing 350 to 450 computer centers housed in carefully selected sites of 12,000 to 15,000 square feet each with adequate security, cooling and wiring for electronic data processing. Each district office would require roughly 100 people which is less than any other option currently under consideration. At the same time, however, the Census Bureau would have to identify, train, and retain a very large number of managers with the ability to run effectively a combined collection and processing office. These are some of the characteristics and requirements of fully decentralized collection and processing for 1990 census-taking.

### Pros and Cons

The most appealing aspects of full decentralization follow:

#### Staffing

- We would not need as many personnel in one location and could get better employees, such as hiring a higher percentage of keyers who are good typists.
- Our wages would be more competitive in the cities where these offices would be located in, than in the cities where more centralized offices would be in. (For the 1986 test, our wages are much more competitive in East Central Mississippi than in Central Los Angeles County.)

#### Workload

- Processing offices would be more manageable; thus, less likely to bottleneck and grind to a halt. The IRS's main problem was that their new systems could not handle the massive loads.

#### Coordination with field activities

- Processing work closer to field collection would improve turnaround time on field assignments. Effectively, we would have more time to actually process data and prepare field assignments.

#### Contingency plans

- In the event of a disaster, work would be distributed to the nearby offices with the greatest ability to absorb it. Since each office would be small, a fire destroying questionnaires would have a much smaller impact than in any other scenarios.

#### Storage of forms

- Because the workloads are small, it is easier to sort forms by geo (CBNA) if clerical access becomes necessary. Clerical access would be necessary if our searching and matching routines have unexpected difficulty, or for some reason we are unable to key names.

The corresponding disadvantages consist of potential obstacles in the following areas:

#### Space

- We would need secure and air-conditioned space in 400+ locations. The automated equipment we will have in the district office is somewhat sensitive to heat and humidity.

#### Technical staffing

- Although the level of technical expertise needed is low, these people may be difficult to hire in some locations.

#### Installation and maintenance

- Traditionally, the Census Bureau has installed and maintained its own equipment. We would have to rely on private contractors to install and maintain equipment.

#### Headquarters control

- Headquarters would not be able to keep as tight a reign on processing operations if they are decentralized. Quality, however, would be relatively consistent as software will drive much of quality and no one office would be large enough to lower the quality of the Nation's data.

#### Software updates

- Software updates would have to be transmitted by telephone and would be more difficult to ensure than in a more centralized processing system.

#### Use of equipment

- This approach would require more backup equipment. In addition, we would need to do a better job predicting workloads than were needed in previous censuses. Once equipment had been installed in one district office, it would be difficult to adjust the equipment profile.

#### Sparsely settled areas

- Staffing may be difficult in some of the more sparsely populated district offices with no significant SMSAs (primarily areas that were conventional in 1980).

#### Regional Decentralization

In this option, processing would be separated from collection. Five to twenty regional offices would serve the district collection offices. The ideal number in this scenario may be 12 to achieve a match with our permanent field office structure. The number of district offices served by a processing office would be determined largely on population density and management's ability to control effectively 15 to 100 offices through one processing office – even when they parallel regional field offices.

This approach keeps down the number of offices for Headquarters to manage and support and enables us to use more labor saving equipment; but at the potential risk of making individual offices difficult to staff.

In this scenario, the processing office receives all mail returns, keys surnames needed for non-response followup, captures data, prepares the failed edit followup cases for the district offices, keys names, matches for coverage improvement operations, and prepares listings for all personal visit followup operations.

Data conversion in the regional processing offices would be done with the Census Bureau's FOSDIC equipment or with a specially designed Optical Mark Reader (OMR).

The primary communications needs are: sending 1.2 to 2.7 million questionnaires back to the

district offices for telephone failed edit followup, 33,000 to 75,000 pages of nonresponse followup listings, and a similar number of edit followup listing pages. The processing office would be receiving 3.5 to 8 million forms from the district offices on a flow basis (nonresponse, edit followup, and vacant/delete).

In this option, field support would have to be carefully coordinated since some district offices might be 2 to 4 hours' drive from the processing offices. Since much of the data that flow between these points are subject to title 13 confidentiality restrictions, it is not likely that telecommunications would solve this problem for us.

### Pros and Cons

The potential benefits of a regional structure follow:

#### Cost

- Scanning approaches – especially scanning both long and short forms – cost the least. Scanning all data would cost less than other conversion options.

#### Headquarters control

- This approach would be the easiest for Headquarters to control due to the low number of offices.

#### Technical support

- The small number of offices would enable the Census Bureau professionals to maintain the equipment.

#### Operational consistency

- In this approach, the regional processing centers would be sized to utilize similar hardware, software, and operating procedures. Management techniques would be interchangeable and quality assurance should be easier to perform than in a hybrid or diverse structure.

However, some disadvantages of this option deserve attention also. They are:

#### Coordination with field activities

- Processing office managers may not be able to develop working relationships with field personnel below the regional office level. Field and processing management systems may not be parallel, meaning that processing offices will cross regional boundaries and add to the coordination problem.

#### Staffing

- The larger offices would be extremely difficult to staff, as they would be much larger than the size of the 1980 processing offices.

#### Technology not proven for this use

- No existing OMR meets our needs. Machines are planned, but not in production. No complete test of using FOSDIC (Film Optical Sensing Device for Input to Computers) without an upfront edit will have been done before the data capture methodology decision must be made. The 1986 test will use 1980 cameras which do not have the resolution needed for 1990 cameras.

#### Space

- The space needed for an average office would cover several football fields. Since this space must have large power supplies and excellent air-conditioning systems, we would have difficulty locating appropriate space.

#### Central control

- Controlling the flow of paper would be very difficult due to the sheer volume of paper moving within the processing office and between the processing office and the district offices. The probability would be high that materials occasionally would go to the wrong district office, delaying field work 2 to 5 days.

#### Distant to district offices

- Some district offices will be over a day's drive from the processing offices. The start of followup operations may be delayed because of long turnaround and delivery times.

The following section deals with the strategic considerations that await our further analysis and resolution before 1986, when all 1990 processing decisions will be made.

### Strategic Considerations

The options just discussed for processing 1990 census data are being tried out also through formal field tests in 1985 and 1986 to simulate 1990 census collection and processing conditions. The investment being made by the Census Bureau to find out more about remote processing centers

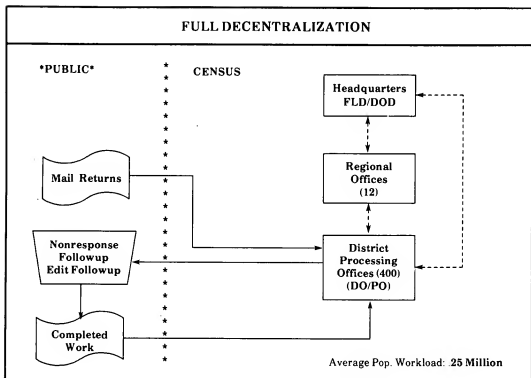
(1986 Census of Central Los Angeles County) and combined collection/processing (1986 Census of East Central Mississippi) is expected to yield a useful return in the form of information for decision-making.

The results of our tests should help us in two ways. First, our operating models will be verified, and adjustments to our current assumptions about production rates, maintenance costs, training difficulties, and many others can be made. Second, and more important, we can assess the management control requirements of decentralization. This aspect of running several processing centers may be the most ominous of all. It brings to mind issues of recruiting 20 to 50 senior managers, locating suitable space, deposing of hardware, establishing fiscal controls, and monitoring the daily activities of a nationwide network of processing centers.

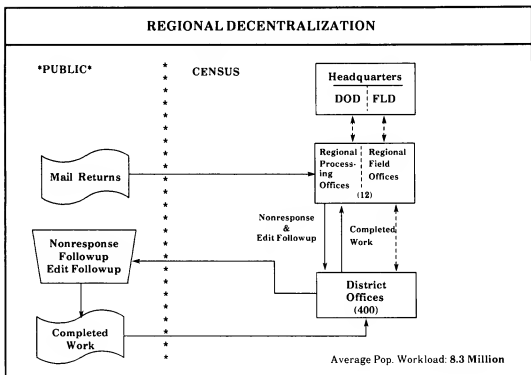
Since the decennial census is a project that emerges, surges, and then subsides, there is no established structure within the Census Bureau to support a far-flung network of processing centers. Accordingly, the decennial census divisions must lean heavily on the Census Bureau's existing experts in data processing, systems engineering, administrative, personnel, and procurement. Moreover, the demands on these groups are much greater if they must support a decentralized setup than if a 1980 style approach is adopted.

So, while many factors relevant to the pros and cons of decentralization will be handled through rigorous analyses and testing, the factor that may truly determine success or failure will be harder to evaluate. This is why the Census Bureau must carefully construct a management approach to decentralized data capture that lessens risk while enhancing the other primary 1990 census goals.

# Exhibit A. Full Decentralization



# Exhibit B. Regional Decentralization



# ON THE USE OF AUTOMATED CODING AT STATISTICS SWEDEN

LARS LYBERG

STATISTICAL RESEARCH UNIT  
STATISTICS SWEDEN

## Introduction

### 1. The coding operation and the characteristics of the control problem

Examples of data-processing operations in a survey are editing, coding, key punching and tabulation. Consider a collective of objects ("elements") of some kind and a set of mutually disjoint categories. Each element belongs to one and only one of these categories. Coding denotes the act of assigning the elements into these categories.

In practice the coding is based upon access to verbal information about the elements of the population or sample under study. This information is usually obtained on schedules in the data collection operation and is entered either by the respondents themselves or by interviewers or enumerators. Unlike certain other kinds of information (numerical data on household expenditures, for instance), verbal information cannot be processed immediately into statistical tables. It must first be coded into different categories where each category is labeled with, for instance, a number. These numbers are called code numbers and the key to these code numbers is called the code. (Naturally, numerical data also may be subject to coding; thus, in a census of businesses the objects enumerated can be assigned to categories defined with respect to, e.g., total turnover.)

The term "coding" is admittedly ambiguous. Attempts have been made to replace it by the term "classification"; this term may be better than coding but it has certain disadvantages. Throughout this article the term "coding" is used since it is the one most frequently used in the literature. Some other terms in this area are ambiguous as well. For instance, what in this article is called "code number", is in the literature often referred to as "code", and what here is called "code" is often referred to as "code list", "coding standard", "lexicon", or "nomenclature". Still another ambiguity concerns what is to be coded. In the definition above it was postulated that a given element belonged to a certain category. In the

literature, coding is often described as an operation in which the verbal descriptions of the responses are coded rather than the elements themselves. This common way of describing the situation is easily understood since in many surveys each element is coded with respect to more than one variable.

The coding operation has three components:

- (1) Each element in, for instance, a population is to be coded with respect to a specific variable by means of verbal descriptions.
- (2) There exists a code for this variable, i.e. a set of code numbers in which each code number denotes a specific category of the variable under study.
- (3) There is a coding function relating (1) and (2), i.e. a set of coding instructions relating verbal descriptions with code numbers.

Coding is a major operation in such statistical studies as censuses of population, censuses of business and labor force surveys. Examples of variables are occupation, industry, education and status.

The problems with coding are of different kinds. As with most other survey operations, coding is susceptible to errors. The errors occur because the coding function is not always properly applied and because either the coding function itself or the code is improper. In fact, in some statistical studies, coding is the most error-prone operation next to data collection. For some variables, error frequencies at the 10% level are not unusual. Another problem is that coding is difficult to control. Accurate coding requires a lot of judgement on the part of the coder, and it can be extremely hard to decide upon the correct code number. Even experienced coders display a great deal of variation in their coding. Thus, there are problems in finding efficient designs for controlling the coding operation. A third problem is that many coding operations are difficult to administer.

Coding has a tendency to become time-consuming and costly: for instance, in the 1970 Swedish Census of Population, carrying out the coding took more than 300 man-years. In many countries, coders in large-scale operations must be hired on a temporary basis, and the consequences for maintaining good quality are obvious. There are even reasons to believe that in the future it might be difficult to obtain even temporary coders for this kind of relatively monotonous work. So there is certainly room for new ideas on the effectiveness of the coding operation.

An overview of the problems with control of coding is given in Lyberg (1981).

## 2. Coding errors in Sweden

### 2.1 The 1965 Swedish Census of Population

In 1967, an evaluation study of coding errors in the 1965 Swedish Census of Population was conducted (see Lyberg (n.d.) and Dalenius and Lyberg (n.d.)). From a population of census material comprising about 70% of the 1965 population a two-stage sample of verified census schedules was selected. The population was partitioned into four strata, and four subsamples were obtained. The evaluation study was confined to the following variables:

- (1) Relationship to head of household
- (2) Type of employment
- (3) Status
- (4) Industry

The codes used for variables (1) to (3) were one-digit codes; the code used for "industry" was a three-digit code.

Since we were dealing with four variables and four subsamples we obtained 16 different estimates of error rates. These are given in Table 1.

Subsamples 1 and 3 consisted of totally verified schedules, and subsamples 2 and 4 consisted of sample verified schedules. Most of the total verification was done by still inexperienced production coders; this explains the differences in error rates between total and sampling verification.

### 2.2 The 1970 Swedish Census of Population

In the 1970 Swedish Census of Population, the number of variables to be coded increased over that in 1965. For evaluation purposes a sample was drawn from the population of census schedules. A pool of expert coders was used to generate a set of "true" evaluation code numbers for each schedule in this sample. These code numbers were compared with the production code numbers after verification, and this led to estimates of error rates for the different variables on economic activity. These variables were:

- (1) Relationship to head of household
- (2) Type of activity
- (3) Occupation
- (4) Status
- (5) Industry
- (6) Place of work
- (7) Type of conveyance to place of work
- (8) Number of hours at work

Estimates of error rates for these variables are given in Table 2.

**TABLE 1. Estimates of Error Rates (%) in Production Coding in the 1965 Swedish Census of Population**

Subsample	Variable			
	(1)	(2)	(3)	(4)
1	1.6	2.7	1.0	14.5
2	1.4	1.6	.6	8.2
3	1.5	3.0	1.3	14.5
4	.7	1.3	1.2	8.7

**TABLE 2. Estimated Error Rates in Coding Economic Activity in the 1970 Swedish Census of Population**

Variable	Code	Per cent error rate (total population)
(1)	1-digit	4.3
(2)	1-digit	4.7
(3)	3-digit	13.5
(4)	1-digit	3.7
(5)	4-digit	9.9
(6)	1-digit	8.9
(7)	1-digit	11.5
(8)	1-digit	4.4

The error rates for the variables (1), (6) and (7) are probably overestimated, since the code numbers were processed by an optical character recognition machine and we have reason to believe that technical errors in this phase had a minor effect on the error rates for those variables.

The table shows that the multi-digit variables are difficult to code. But also the one-digit variables, a priori considered easily coded, are erroneously coded relatively often. One reason could be that the coding situation is too complex for one coder, i.e. each coder has more variables to manage than he/she can handle.

### **2.3 The 1975 Swedish Census of Population**

The number of variables was smaller in the 1975 Census of Population than in the 1970 Census. Evaluation studies show that the error rates also were smaller in this census than in the 1970 Census. The following variables were studied:

- (1) Relationship to head of household
- (2) Type of activity
- (3) Occupation
- (4) Status
- (5) Industry
- (6) Type of employment
- (7) Type of conveyance to place of work

All of these are one-digit variables except for (3) and (5). In Table 3 estimated error rates are given for these variables.

The results given in this table differ strikingly from those obtained in the 1970 evaluation study. The error rates have dropped for every variable and the fact that the one-digit variables now really seem to be easily coded is most encouraging. The occupation error rate

**TABLE 3. Estimated Error Rates in Coding Economic Activity in the 1975 Swedish Census of Population**

Variable	Code	Per cent error rate (total population)
(1)	1-digit	.6
(2)	1-digit	.6
(3)	3-digit	7.8
(4)	1-digit	.5
(5)	4-digit	3.5
(6)	1-digit	1.0
(7)	1-digit	.5

of almost 8% is still very serious, but compared to the 13.5% rate in 1970 it is a good result. Even better is the estimate for industry.

## **2.4 Some other studies of error rates at Statistics Sweden**

Most of the coding studies at Statistics Sweden have been carried out within the censuses. This is rather natural since the coding is a very extensive operation in a census. During the last decade interest in coding errors has grown and as a result some evaluation studies have been carried out in other surveys as well. Here some estimates of coding errors from such studies are given.

In Olofsson (1976) an industry error rate of 5.7% is noted in the 1974 Labour Force Survey. Occupation in the same survey had an error rate of 6.2%. In Harvig (1973b) an 11% error rate in occupation coding is estimated for coding data for university graduates. In Harvig (1973a) a 3.2% error rate is estimated when coding underlying causes of death. In Lyberg et al. (1973) an 8% error rate is estimated when coding teachers' education. In this case the 95% confidence interval was 5.9% - 10.3%.

Extensive reviews of studies of error rates in industry and occupation coding are given in Lyberg (1983).

## **3. The need for control**

It is imperative in most statistical series that coding control is made part of the overall program for producing the statistics. However, knowledge of the error rate is not enough if we want to be far-sighted. We need to know about the error structure, the reliability of the coding process, different types of errors, the seriousness of different errors and the effects of errors, in order to take suitable corrective measures with respect to the code or the coder.

Several control options are available:

Firstly, by means of, say, the U.S. Bureau of the Census' survey model, it is possible to dissect the coding error in a given coding operation. Such a model can also help strike an appropriate balance between various control efforts with respect to all survey operations. (See Bailar and Dalenius (1969).)

Secondly, manual coding is rather well suited for the application of statistical quality control schemes as originally developed for industrial applications. (See Minton (1969).)

Thirdly, there are certain control schemes designed specifically for coding. Two such main schemes are called dependent and independent verification. (See Lyberg (1981).)

Fourthly, evaluation of coding results provides a basis for the allocation of quality control efforts. We have already given examples of results from different evaluation studies. The results of such studies give suggestions concerning the size and emphasis of the necessary quality control program.

Fifthly, it appears inevitable to focus on the very basis of manual coding and to consider the possibilities offered by access to a computer of developing a basically new approach. This idea is, of course, not in principle new: for instance, at the U.S. Bureau of the Census, geographic coding has been conducted by means of computers since 1963. What is new is the suggestion in that agency that the computer be used extensively in the coding of such complex variables as occupation and industry. This suggestion may be viewed as a natural extension of earlier uses of computers in the editing operations.

The remaining part of this paper describes the Swedish efforts in this specific field of automation.

## **Automated Coding - An Overview**

### **4. A bird's-eye view of automated coding**

In automated coding we distinguish four operations:

- (i) construction of a computer-stored dictionary;
- (ii) entering element descriptions into the computer;
- (iii) matching and coding;
- (iv) evaluation.

#### **4.1 Construction of a computer-stored dictionary**

In automated coding a dictionary stored in the computer takes the place of the coding instructions and the nomenclature used in manual coding. Obviously, the construction of such a dictionary is a very important task. The construction

work could be carried out manually but, when dealing with complex multi-digit variables, using the computer seems to be a better alternative. The resulting dictionary should consist of a number of verbal descriptions with associated code numbers. The descriptions could be a sample from the population to be coded or a sample from an earlier survey of the same kind. Of course an important problem is the size of the sample underlying the dictionary construction. Whether the dictionary is constructed manually or by computer, the code numbers appearing in it should ideally be those assigned by the best of the available coders and controlled by means of efficient independent verification procedures.

#### **4.2 Entering element descriptions into the computer**

Verbal descriptions are to be entered into the computer. One possible method is to punch the descriptions in a more or less free format. However, this method has some serious drawbacks: first it consumes a lot of "space", and second, the errors involved in large-scale keypunching of alphabetic information are relatively unknown; moreover, such keypunching is rather costly.

A better alternative would be to have the verbal information directly available for optical character recognition. Unfortunately the recognition of hand-written letters is not yet sufficiently developed for this purpose.

There are reasons to believe that at present the entering of verbal descriptions into the computer is the most important practical problem in designing systems for automated coding.

#### **4.3 Matching and coding**

Each element description now put into the computer is compared with the list of occupation descriptions in the dictionary. If an element description agrees with an occupation description (is a "match"), it is assigned the corresponding code number; otherwise, it is referred to manual coding.

In an automated coding system we will obtain exact matching for a fraction of all elements only. A primary task in developing such a system is to design

criteria for the degree of similarity between input words and dictionary words necessary for them to be considered to match.

#### **4.4 Evaluation**

The system must include continuing evaluation studies. Such studies aim at:

- (i) controlling the quality of computerized coding;
- (ii) improving the dictionary; and
- (iii) controlling the cost.

Whether automated coding is economical or not is a question to be answered by the evaluation. Are the referred cases more difficult to code than those taken care of by the computer? Does the dictionary need improvement? These and other questions are to be resolved by evaluation.

#### **5. The dictionary**

There are two general kinds of algorithms for automated coding: weighting algorithms and dictionary algorithms. Weighting algorithms assign weights to each word-code combination using information from a basic file: when a new record is to be coded, the program chooses the code number which is assigned the highest weight for the specific record word. Dictionary algorithms look in a dictionary for words or word strings which imply specific code numbers: when a new record is to be coded, the program determines whether the word or word string matches any word in the dictionary. If no match occurs, the record is rejected and referred to manual coding.

At the U.S. Bureau of the Census a number of different algorithms have been developed and investigated during the last decades. In some straightforward applications like the geographic coding, automated coding has been quite successful. Recent efforts deal mainly with the more complex coding of occupation and industry. Four algorithms are described in Lakatos (1977a, b). Two of them, the O'Reagan and the Corbett algorithms, use dictionary methods. The remaining two, the IMP and the INT algorithms, use the weighting method. The INT algorithm is due to Rodger Knaus, and is further described in Knaus (n.d., 1978a, b, 1979, 1983). Current development work at the U.S. Bureau of the Census is described in Appel and Hellerman (1983) and Appel and Scopp (1985).

At Statistics Sweden we have worked with the dictionary approach only. Thus, we have nothing to add with respect to other algorithms.

Thus, the computer-stored dictionary is a parallel to the dictionary and the coding instructions used in manual coding. In order to create such a dictionary a number of operations must be carried out:

- (i) choice of basic material;
- (ii) sampling a basic file from the basic material;
- (iii) expert coding of the basic file;
- (iv) establishing inclusion criteria for dictionary records;
- (v) construction of a preliminary dictionary;
- (vi) testing and completing the preliminary dictionary.

### 5.1 Choice of a basic material

The most suitable basic material is the set of filled-out forms in the survey under study. To use this is rarely possible – time is not on our side. Instead the basic material must often consist of:

- (i) material from an earlier survey of the same kind; or
- (ii) material from a pilot survey; or
- (iii) material from another kind of survey in which the same variable was included.

It should be pointed out, though, that basic material of the desirable kind implied above could be efficiently used when revising a dictionary that has been used in production for a while.

It is important that the basic material be up to date. Structural changes occur in the population; e.g., entry and exit of industry and occupation denominations occur frequently. Also, it is possible that the respondent-reporting pattern changes over a period of time. One example could be the following. In the 1965 Swedish Census of Population, cleaners used to describe their occupation as "cleaner". In the 1970 census, a new term, "local keeper", was used by some cleaners. That term had not existed in 1965 and as a consequence was not represented in the basic material. The result was that the dictionary based on the 1965 census material was not able to code 1970 census individuals describing their occupation as "local keeper".

Basic material as in (iii) should only be used in exceptional cases, since the reporting pattern for a certain variable could differ substantially between different surveys due to different modes of data collection.

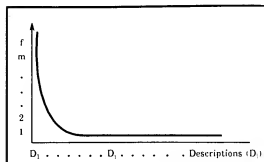
### 5.2 Sampling a basic file from the basic material

From the basic material we must sample a number of records in order to construct a dictionary. The sampling of records could be carried out in different ways, for instance:

- a simple random sample,
- a controlled random sample, or
- a subjective sample.

With the first approach, descriptions with low frequencies have a small probability of being included in the file. This is generally not a negative consequence. Therefore, in almost all of the experiments and applications conducted at our agency we have used simple random sampling. The sample size is a problem, irrespective of the kind of approach we use, since each description should be coded by "experts".

In some of the experiments with automated coding conducted at the U.S. Bureau of the Census, a very large initial random sample of records was chosen: sample sizes of about 100,000 records have been used. In the experiments at Statistics Sweden, the basic file has consisted of at most 14,000 records. Despite that, evaluation studies show comparable results. Possible explanations are that a few code numbers and a few dictionary descriptions are, for many variables, sufficient to code a large portion of the records and that the Swedish language is less complex (at least in this context) than English. A typical frequency diagram for unique descriptions is the following:



The typical diagram has a very straggling tail provided that the descriptions are ordered with respect to the frequencies with which they occur. In fact, in some applications, many unique descriptions occur only once or twice. In O'Reagan (1972) a closer look revealed that, for one variable, 7% of the code numbers could handle 50% of the records. Thus, by means of a rather small initial sample, it is usually possible to get a decent dictionary. Our experiences show that vast increases of the basic file (once the "decent" criterion is fulfilled) do not add much with respect to coding degree. An efficient strategy seems to be to concentrate one's effort on the most frequently used categories and accept manual coding of most of the remaining part.

### 5.3 Expert coding of the basic file

In order to construct a good dictionary the basic file has to be coded with high quality, and for this work we have to use the best coders available. Since even "expert" coding is susceptible to errors, the expert coding of the basic file must be carried out in conjunction with a control operation.

### 5.4 Establishing inclusion criteria for dictionary records

The verbal descriptions in an expert-coded basic file can be classified into different categories:

- (a) descriptions of high frequency which all point at some specific code number;
- (b) descriptions of low frequency which all point at some specific code number;
- (c) descriptions of high or low frequency with which different code numbers are associated.

In principle, all descriptions pointing at some specific code number should be included in the dictionary. Whether this can be done in practice depends on how large a dictionary we can accept. This in turn is a function of the searching time of the matching program. If the searching time is independent of the size of the dictionary and if having an extensive dictionary does not imply lots of manual administrative work, then all descriptions pointing at specific code numbers

should be included. Otherwise, we must define what is meant by "high frequency". This decision depends on sample size and number of categories of the code, among other things; for instance, a small enough basic sample generates no highly frequent descriptions at all. A simple piece of advice is to have a low value of the concept "high frequency"  $f$  say  $f \geq 3$ , since it is always easier to remove than to add descriptions to the dictionary.

Descriptions belonging to category (c) should not be part of the dictionary. There are possible exceptions, though. If, for a vast majority of the cases, a high-frequency description is associated with a specific code number, then an inclusion might be considered. Of course, if such a description is included we end up with deliberately built-in erroneous classifications. Even if such error rates are admittedly small, it is probably better to change the nomenclature so that the coding of this specific description becomes unambiguous in the first place.

### 5.5 Constructing, testing, and completing the preliminary dictionary

A dictionary can be constructed by man or by computer. Presumably, a combination of the two is the most effective. In our first experiments at Statistics Sweden we used manually constructed dictionaries but nowadays we have access to a computer program for dictionary construction.

The manual construction of dictionaries can be characterized as trial and error. At Statistics Sweden we have worked with two lists: list No. 1 is the expert-coded file sorted with respect to code number, and list No. 2 is the same file sorted alphabetically. These lists form the basis for the construction. List No. 1 is used to get some hints about the structure of the verbal descriptions sorted under a specific code number. We choose a frequency limit  $f$  for defining "high-frequency" descriptions. All descriptions occurring  $f$  or more times are stored in the preliminary version of the primary dictionary which is scanned first in automated coding. We call this dictionary PLEX.

In order to increase the coding degree we must include some variants of the high-frequency descriptions already stored.

One possibility is to recognize discriminating word strings. In the ideal situation one such string represents many variants of a certain description. Thus, after storing the high-frequency descriptions, we start looking for discriminating word strings. These strings (or rather, parts of words) are stored in a secondary dictionary. This secondary dictionary, called SLEX, is scanned if PLEX fails to code.

List No. 2 is used as a check. Has a preliminary description stored in PLEX been assigned any other code number except for the specific one under study? It is common that a certain description can be associated with different code numbers depending on the code, the coding instructions, and the auxiliary information used by the coders. The alphabetic list helps us identify such descriptions. When they are identified they can be omitted from the preliminary PLEX. The same goes for the associated word strings in the preliminary SLEX. However, as mentioned above, if we deliberately permit a certain degree of erroneous coding, some of these ambiguous descriptions may remain. The probability for such a misclassification should be small, though.

Often a number of highly frequent descriptions are omitted because of their lack of unambiguousness. Then, one might reconsider the inclusion of low frequent but unambiguous descriptions in PLEX. Another approach is the possibility to transform some ambiguous descriptions into unambiguous ones by means of auxiliary information.

The word strings in SLEX should be common to several descriptions or be parts of special highly frequent descriptions. We have to be sure that SLEX words do not fit PLEX descriptions for other code numbers. SLEX can never be allowed to expand because of the difficulty to keep up its accuracy. The main problem with SLEX is that we do not know in advance how it behaves when new records are coded.

The manual work described above (or similar manual procedures) can to an important extent be carried out by a computer. Two approaches developed by the U.S. Bureau of the Census are presented in O'Reagan (1972) (O'Reagan's algorithm) and in Corbett (1972) and Owens (1975) (Corbett's algorithm).

The computerized dictionary construction system at Statistics Sweden generates a dictionary with two chapters, PLEX and SLEX. PLEX contains unequivocal descriptions and is scanned first. SLEX contains discriminating word strings that fit several different input descriptions. As a consequence, SLEX is not as accurate as PLEX and it is scanned only if PLEX fails to code. Our experience shows that it is rather easy to construct a PLEX manually, but that manual SLEX construction is much harder to manage. Thus, we have made a program for computerized construction of SLEX. (As a consequence, a computerized PLEX is obtained as a simple special case.)

We have tried a few different versions of the program. The present version, a package called AUTOCOD, is described in Bäcklund (1978). All programs are written in PL1. AUTOCOD contains routines for:

- the creation of computer-stored dictionaries (PCLEXK)
- the coding of descriptions (PCAUTOK)
- the updating of dictionaries (PCLEXUP)
- the evaluation of dictionaries (PCLEXT)

PCLEXK creates a PLEX and a SLEX. The procedure involves three steps. The program LEXLADD creates space for a possible SLEX. LEXKONS creates PLEX and SLEX. For each PLEX description, say, a six-character abbreviation starting with the first character is tested for inclusion in SLEX. If that abbreviation fits another PLEX description it is rejected and a new abbreviation is created starting with the second character of the PLEX description. The procedure is repeated at most six times; if no valid abbreviation is obtained, the procedure goes on to the next PLEX description. Finally, LEXLIST lists the dictionaries by means of EASYLIST. Parameters that can be varied include:

- possible use of a list of prefixes which, when making a dictionary of, say, goods, removes such word strings as pounds, roll, and pairs;
- minimum frequency  $f_0$  (the dictionary inclusion criterion);
- tolerated degree of equivocality;
- minimum length of words in SLEX.

PCAUTOK codes new records by means of PLEX and SLEX. PCLEXUP is used when we want descriptions to be removed from or added to an existing dictionary. PCLEXT is used to evaluate a dictionary when we have access to a material with manual code numbers assigned.

PLEX and SLEX can be updated simultaneously or separately.

## 6. The use of auxiliary information

In manual coding we often use not only the verbal descriptions for the variable to be coded but also different kinds of auxiliary information. Typically, this information consists of descriptions on some related variables. For instance, information on education or industry is sometimes used as auxiliary information when coding occupation.

Of course, auxiliary information can be used in automated coding as well. The necessary conditions are that the auxiliary information is given together with the record descriptions to be coded and that the auxiliary information is also present in the dictionary. Storing auxiliary information in the dictionary and designing the computer programs to allow this kind of matching and coding present no serious problem per se. Especially, the auxiliary information can be used efficiently if the coding is conducted in two steps; results obtained in the first-step coding can be used as auxiliary information in the second-step coding. If the first-step variables are coded manually, the resulting code numbers can be punched together with the verbal descriptions of the variables to be coded in the second step. Since punching of verbal descriptions is a time-consuming operation, a faster publication of first-step results is made possible. The time saving in an extensive investigation such as a census of population may be considerable; especially this is the case if the second-step variables are difficult to code. An example of such a case is the occupation coding in the 1980 Census of Population.

## 7. Evaluation and control

A final and necessary step in an automated system is evaluation and control. Its primary goal is to maintain the pre-specified level of accuracy.

The coding degree,  $p$ , and the proportion correctly coded,  $q$ , are the main characteristics studied for control and evaluation purposes.

If  $N$  is the number of elements entered into the computer and  $n$  is the number actually coded, then  $p = n/N$ . If  $m$  out of the  $n$  coded elements are correctly coded, then  $q = m/n$ . When evaluating an automated coding procedure,  $p$  must be judged together with  $q$ . Obviously, it is more important to have a large value of  $q$  than a large value of  $p$ . When comparing different results, the product  $pq$  might be helpful. Unfortunately, this measure must be used with great care. For instance, the combination  $p = .5$ ,  $q = .9$  is much better than  $p = .9$ ,  $q = .5$ . One should strive primarily for a  $q$ -value as high as possible. After that, one can concentrate on increasing  $p$ . This proportion could be increased until  $q$  starts to decrease. It is even possible to increase  $p$  at the price of a reduction in  $q$ , but then the monetary payoff must clearly outweigh the loss in quality.

The cost for manual coding of the proportion  $1-p$  plays an important role in calculating the costs of the entire coding operation, including both automated and manual steps. The descriptions which the computer is unable to code can be more complex than those it does code. Besides, there is a relatively higher fixed cost associated with the manual coding of the proportion  $1-p$  compared with manual coding of all elements and, furthermore, all manual code numbers must be keypunched. These costs must be considered when evaluating automated coding. However, recent experiences show that in the census application a good bit of the  $1-p$  may be coded without access to the questionnaire which makes the process faster than conventional manual coding.

A secondary goal of the evaluation and control operation is to gather information that can be used as a basis for changes in the dictionaries and the matching programs. Samples can continuously be drawn from the production and coded by skilled verifiers according to some suitable scheme (for instance, independent verification). Thus,  $q$  can be estimated continuously. If  $q$  does not meet quality standards, the sample under study must be analyzed. What types of descriptions have been erroneously coded and which have not been coded at all? Is the sample extreme in some sense? Are special sections of the code difficult for the computer? Of course we could try to answer these and other questions even if the actual  $q$ -value meets the pre-specified standards. However, adjustments should generally be carried out only when the process is out of control or in danger of becoming so.

## Experiments

Over the years, Statistics Sweden has experimented a lot with automated coding. Eventually, the experimental results became so convincing that it was decided to implement the technique in ongoing surveys. In this section we will present some general experiences from the early experiments that might be informative to other researchers.

### 8. Industry

The very first experiments with automated coding at Statistics Sweden concerned the industry variable. These experiments were not very successful. The q-values did not exceed .83, and in one experiment it was as low as .69. During this first phase we thought that a system for automated coding had to be rather complicated and even sophisticated. We were convinced that relying on exact matching only would be highly unsatisfactory. The response pattern, we thought, is so complex that the coding degree obtained by exact matching rules would be much too low to pay off. Therefore, we tried some special matching rules. We experimented with measures of the "distance" between respondent descriptions and dictionary descriptions. We also tried to apply Spearman's rank correlation coefficient as a measure of the similarity between these two kinds of descriptions. More recent experiences have shown that, depending on the level of ambition, this might not be necessary.

After these initial experiments we have not been working with the industry variable very much. In fact, almost all the "trial and error" work, in our opinion the very essence in developing methods for automated coding, is still waiting to be done for this variable. It could even be argued that, most of the time, industry descriptions without access to auxiliary information are more or less useless to manual coders as well. Thus, descriptions of industry only are unsuitable for automated coding of that variable. We were anxious to show the sponsors some results and we started to cast our eyes in another direction, towards the occupation variable.

### 9. Occupation

Most of our experiments have concerned the occupation variable. In Table 4 the main experiments are summarized.

As can be seen from the table, the q-values range from .85 to .95 which we considered gratifying. During this experimental period the program for automated dictionary construction was refined. The different sophisticated matching rules could be used on request, dictionary sizes varied between 900 and 11,000, and we found that SLEX should be used with great care. The quality never increases by means of SLEX. The coding degree increases, of course, but always to the price of decreased quality. For instance, in experiment 7, the quality increased from .84 to .92 when using PLEX only while the associated coding degree decreased from .84 to .69. It is interesting to note that minor changes in PLEX have no effect on the coding degree and the quality. However, there exists an alphabetic list of approximately 11,000 occupation descriptions. This list, used by manual coders, is not based on knowledge of the empirical response patterns. We were anxious to know what would happen if that list, already available on tape, was used. We did that in experiments 8 and 9 and obtained the coding degrees 40.2% and 36.0% respectively. Thus a desk product such as this list cannot compete with a PLEX based on empirical response patterns, although the latter in this study consisted of 1,637 descriptions only. When we merged our computerized PLEX with the alphabetic list and used them together as an extended PLEX, the coding degree increased to approximately 75%. Thus, such a combination could be useful.

### 10. Goods

Goods or purchase is a main variable in household expenditure surveys. Coding of this variable is relatively simple compared with coding, say, occupation. The normal case is not complicated. Observed coding error frequencies from different household expenditure surveys support this assumption. Experiments with this variable, using PLEX only, resulted in q-values around .995 and coding degrees around .68. The fact that this coding is uncomplicated but costly and time-consuming made it an excellent automation prospect.

### Applications

Automated coding has been applied in some regular productions at Statistics Sweden. The very first application was the coding of goods in the 1978 Household Expenditure Survey. After

**TABLE 4. Experiments With Automated Coding of Occupation**

Experiment	Type of dictionary	Survey	Coding degree (%)	Quality or agreement rate (%)
1	Manual	1965 Census	62	95
2	Manual	1970 Census	66	92
3	Manual	1970 Census	74	84
4	Manual	1970 Census	80	90
5	Manual	Labor Force 1974	81	81
6	Computerized (PLEX + SLEX)	1970 Census	69	87
7	Computerized (PLEX + SLEX)	Labor Force 1976	84	85
8	Computerized (PLEX)	Labor Force 1976	69	93
9	Computerized (PLEX)	Labor Force 1976	69	92
10	Computerized and manual combined	Labor Force 1976	74-76	93-94

that, automated coding has been applied in coding occupation and socio-economic classification (SEI) in the 1980 Census of Population and in coding goods in the 1985 Household Expenditure Survey. The procedure is also used in a continuing survey of book loans where authors and book titles are coded, in coding occupation and SEI in the continuing Survey of Living Conditions and in coding occupation in pupil surveys. Other applications are underway.

## **11. Coding goods in the 1978 and 1985 Household Expenditure Surveys (HES)**

### **11.1 Introduction**

In the 1978 HES, approximately 5,900 households were supposed to keep a complete diary (CD) of all goods purchased during a two-week period. The rest of the sample, approximately 7,900 households, was supposed to keep a simplified diary (SD) of goods purchased during a four-week period.

In the 1985 HES, approximately 6,000 households were supposed to keep a diary of goods purchased during a four-week period. The complexity of this diary corresponds to a case somewhere between the 1978 CD and SD. The

survey still strives for a reasonable completeness while all the foodstuffs are assigned the same code number.

The survey designs allow continuous delivery of diaries from the respondents during the survey year. Thus, the material can be processed in cycles, which might be advantageous in a system with automated coding. In 1978 we were not at all convinced that our system should work in a production environment, so it was decided that, to start with, the coding should be carried out by two parallel systems, one manual and one automated. After two months of production, the systems were evaluated in order to decide a preferred system to be used during the remaining 10 months of production, and automated coding passed the test. In the 1985 survey, automated coding was used from the start.

### **11.2 The automated system**

In 1978, the dictionary construction was step-wise. Extensive efforts were made in creating an initial dictionary. After that, continuous revisions were made prior to many cycles. The initial dictionary was based on the dictionary used in

the experiments mentioned in the Experiments section together with a list of all descriptions in the experimental material. Each unique description was coded by HES experts. The construction involved a lot of manual work since the pattern of descriptions had changed during the nine years that had passed since the last 1969 HES from which we had gathered the experimental material. Only a PLEX was constructed, with a 100% unequivocal rate. This initial dictionary consisted of 1,459 descriptions. In the automated coding procedure, uncoded descriptions were listed alphabetically on an optical character recognition form, and code numbers were assigned directly on it. Some of the uncoded descriptions were added to the dictionary in the updating process.

In 1985, the 1978 dictionary (the last version, 4,230 descriptions) was used as a starting point. This manual procedure resulted in a first dictionary consisting of 1,985 PLEX descriptions. When this is written, the 1985 HES is still going on.

### 11.3 Results

In the 1978 survey 33 cycles were run. During this period 17 different versions of PLEX were used; thus, only a few cycles were coded with identical dictionaries. In Table 5 below the dictionary sizes and coding degree for the cycles are given.

The coding degree over all cycles was 65%. As can be seen from the table, the coding degree decreases sharply now and then. This is explained by the fact that CDs are easier to code automatically compared with the SDs and that the proportion of CDs varies between the cycles. As a matter of fact, some cycles contain only one type of diary. An estimate of the coding degree for CD is 70% and for SD, 38%.

The dictionary was modified prior to most of the cycle runs, at least for the major part of the production. As shown in Table 5, the additions have generally outnumbered the removals. These

**TABLE 5. Dictionary Size and Coding Degree for the 33 Cycles in the 1978 HES**

Dictionary version	Cycle	Number of dictionary descriptions	Coding degree (%)
1	1	1,459	56
2	2	1,554	63
3	3	1,760	67
4	4	2,228	66
5	5, 6, 7	2,464	68, 68, 63
6	8	1,632	64
7	9	1,990	53
8	10, 11	2,451	69, 66
9	12	2,866	61
10	13	3,065	68
11	14	3,613	58
12	15, 16	3,752	72, 73
13	17, 18	3,832	39, 70
14	19, 20	4,011	65, 73
15	21, 22	4,229	51, 72
16	23, 24, 25, 26, 27, 28, 29, 30	4,230	64, 67, 62, 67, 72, 65, 67, 50
17	31, 32, 33	4,230	65, 39, 67

modifications did not change the coding degree very much, though. A closer look reveals that many dictionary words are used very seldom or not at all and that relatively few dictionary words can take care of most of the input descriptions. All new uncoded descriptions were gathered in a special file. As soon as a new description occurred in at least three households it was included in the dictionary, provided it could be unequivocally coded to a specific category.

Essentially, the same procedure is currently used in the 1985 HES. In Table 6, the dictionary size and coding degree for the 17 cycles run so far are given.

In coding these HESs it was decided to use PLEX only because of the inefficiency of the SLEX file. We assume that the coding degree goes down 10-15 percentage points when SLEX is dropped. However, for the 65% coded, the coding quality is high with an error rate less than 1%. Special evaluation studies showed that the quality of the coding of the remaining part was very good too: the error rate was around 1%. This rate can by no means compete with the one a SLEX would give. In all, the coding of the 1978 HES was a smooth

operation. The key operators found it less boring to punch verbal information for a change. The cost calculations point to the fact that automated coding was 2-5% cheaper than a conventional manual system. Besides, the system provided some further advantages. Since all descriptions are keypunched, the primary material is better documented than when merely the code number is keyed. Thus, it is possible to give more detailed descriptions of the goods contained in the groups for which estimates are provided. Furthermore, since the dictionary manages to code most straightforward descriptions, the remaining manual coding becomes more interesting to the coder. The coding of the 1985 HES is a smooth operation, too. We have learned from the experiences gained in 1978 and cut down on administrative routines.

As can be seen from the tables above, it does not seem worth the effort to make extensive dictionary revisions after a specific point. Quite soon a rather stable coding degree is obtained which cannot be substantially altered without changing the dictionary construction principle. For the 1978 survey, we note that with the third version already we have obtained a coding degree of 67%. Despite much work and repetitive modifications

**TABLE 6. Dictionary Size and Coding Degree for the 17 Cycles Run so Far (September 1, 1985) in the 1985 HES**

Dictionary version	Cycle	Number of dictionary descriptions	Coding degree (%)
1	1, 2	1,985	81, 78
2	3	2,029	82
3	4, 5	2,063	82, 81
4	6	2,126	82
5	7	2,156	83
6	8	2,176	82
7	9	2,207	81
8	10	2,228	83
9	11, 12, 13	2,272	83, 83, 82
10	14, 15, 16	2,370	83, 80, 81
11	17	2,446	83

after that point, we have at best obtained 73%.

The 1985 survey has a similar pattern so far. The dictionary has not grown as much as it did in the 1978 survey, but the growth that actually has taken place (from 1,985 to 2,446 descriptions) has not affected the coding degree which is very stable in the interval 81- 83%.

The computer coding is very inexpensive. The coding of tens of thousands of descriptions costs less than 50 Canadian dollars. Computer costs for updating dictionaries are about the same. The expensive part is manual preparation and administration. As mentioned, this part is conducted more efficiently in the 1985 survey.

## **12. Automated coding of occupation and socio-economic classification in the 1980 Census of Population**

### **12.1 Introduction**

In the 1980 Census of Population the coding of occupation and socio-economic classification (SEI) is automated. In short, this automated coding means that personal identifications and the occupation descriptions are punched and matched against a computer-stored dictionary. The dictionary contains occupation descriptions with associated occupation and SEI code numbers.

The occupation code used in the census is built upon the "Northern Standard for Classification of Occupations" (NYK) which in turn is built upon the "International Standard for Classification of Occupations" (ISCO). The code contains roughly 280 different three-digit categories. The code for SEI is a two-digit code with 14 different categories.

Here we shall concentrate upon the coding of occupation, since the system was originally constructed for this coding. The coding of SEI was added later on and the system is not "perfect" for coding that variable.

### **12.2 The coding system: an overview**

In Figure 1 a chart of the coding system is shown.

First, the occupation descriptions and the personal identifications on the census questionnaires are keypunched. The punched information from a questionnaire is called a questionnaire record. A questionnaire record may contain one or two individual records. After the keypunching, the questionnaire records are split into individual records and at the same time punched occupation descriptions are edited.

In the editing process, special signs (points, lines, etc.) and prefixes (1st, vice, etc.) are removed and the remaining parts of the occupation description are brought into one sequence.

The punched file is matched against a file containing the economically active population in the census. In this matching we get some unlinked punched records, for example, due to the fact that occupation is punched for an individual who is not economically active. These unlinked punched records are not used henceforth. Punched occupation descriptions will be missing for some economically active individuals. This may be due to the fact that some occupation description on a specific questionnaire is missing. Sometimes the description may be present on the questionnaire but it has been omitted in the keypunching process.

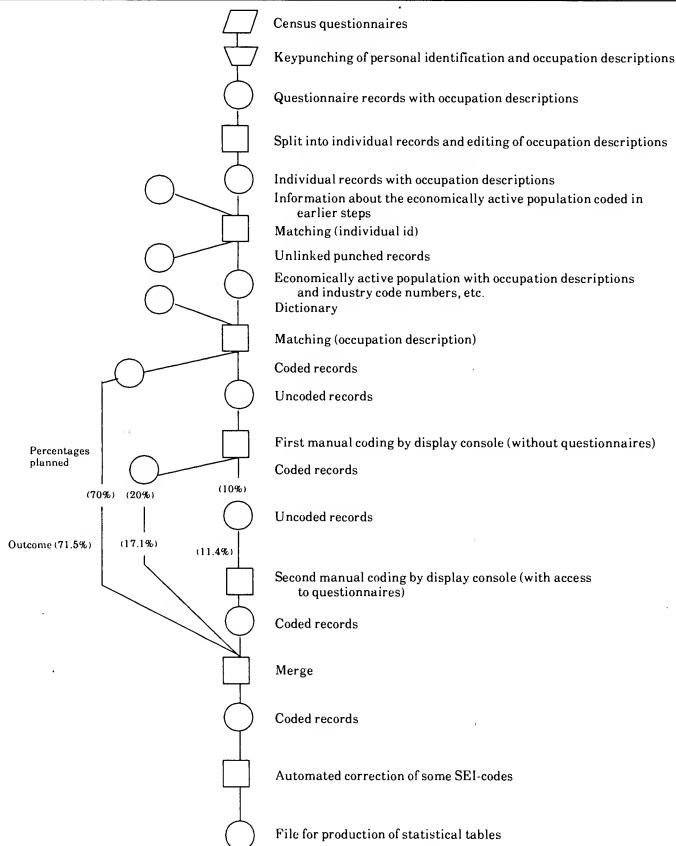
All economically active individuals must of course be coded, at least into some of the "trash" categories designed for situations where the occupation is unknown. In connection with the matching, code numbers for type of activity, industry, institutional classification and so on are obtained from the file of the economically active.

As a result of the matching we get a file which contains among other things:

- personal identification;
- punched and edited occupation description (with the exception mentioned above);
- industry code number;
- institutional classification code number;
- size of establishment.

This file is sorted according to edited occupation descriptions and industry

**Figure 1: Coding of Occupation and SEI in the 1980 Census of Population – An Overview**



code numbers and matched against the computer-stored dictionary. If an edited occupation description is found in the dictionary, then occupation and SEI are coded.

The census occupation dictionary contains both PLEX and SLEX, and it is described in more detail below.

The manual coding is carried out on display consoles in two steps. The first manual coding is carried out without access to the questionnaires. The records which cannot be coded are left "empty" and are coded later on in the second manual coding. In the second step the questionnaires are used. Then, the automatically coded records and the records coded in the first and second steps are merged into one file.

At last some SEI-code numbers are automatically corrected. This correction is made by means of a specific question on the questionnaire, where the variable associated with that question has been coded in an earlier step. This question provides information whether the respondent is an employer or an employee, which is an important aspect of the SEI-code.

## 12.3 The dictionary

The dictionary consists of a PLEX and a SLEX. An excerpt from PLEX is given below.

There must be an exact agreement between an input occupation description including any auxiliary information and a PLEX dictionary description to be considered a "match". PLEX is using industry (NÄRG), institutional classification (SEKT) and size of establishment (S) as auxiliary information.

Since the coding operation was carried out in two steps, we have a most favorable situation for automated coding. First, type of activity, industry and some other variables were manually coded. Then the automated coding of occupation and SEI was carried out. As already pointed out, this results in certain time savings when it comes to publishing results concerning the variables coded in the first step. Besides, two-step coding makes it possible to use the auxiliary information in the automated coding process. We believe that the good result of the automated coding in the 1980 Census (presented below) is, to a large extent, due to the fact that auxiliary information was used.

### ALFABETISKT PRIMÄRLEXIKON

YKOD	SEI		NÄRG	SEKT	S
793	11	TAKARBETARE			
793	11	TAKLÄGGARE			
793	11	TAKMONTÖR			
032	60	TANDL			
032	60	TANDLÄKARE			
044	36	TANDSKÖT			
044	36	TANDSKÖTERSKA			
044	36	TANDSKÖTERSKEELEV			
044	36	TANDSKÖTERSKEPRAKTIKANT			
744	46	TANDTEKNIKER			
032	60	TANDVLÄKARE			
044	12	TANDVÄRDSBITRÄDE			
633	12	TANKBILSCHAFFÖR			
801	11	TAPETFABRIKSARBETARE			
781	21	TAPETSERARE	50		
714	21	TAPETSERARE	*		
781	21	TAPETSÖR	50		
714	21	TAPETSÖR	*		
801	21	TAPETTRYCKARE			
504	11	TAPPARE	2		
821	11	TAPPARE	31		
772	11	TAPPARE	33		
736	11	TAPPARE	37204		
731	11	TAPPARE	37		

When studying the excerpt on the previous page, it is, for instance, seen that the description "TAKLÄGGARE" (Roof builder) always gets the code numbers 793 and 11 for YKOD and SEL, respectively. The description "TAPETSERARE" (Upholsterer) gets the code numbers 781 and 21, respectively, provided the industry code number is 50 (construction). For all other industry code numbers (=\*) the description "TAPETSERARE" gets the code numbers 714 and 21, respectively.

SLEX contains parts of words of the type "ADJUNK" (part of the word ADJUNKT which means something like "assistant master at secondary school"). The purpose is that such a part shall fit many variants of a specific occupation description. "ADJUNK", for example, fits all variants in the example below.

Of course, it happens easily that a certain SLEX-word fits the "wrong" occupation description. It is difficult to avoid such mistakes when building SLEX. One way to reduce the coding errors due to a "coarse" SLEX is to use auxiliary information, for example, industry code numbers. In the example below, "ADJUNK" in industry 931 (education) is coded with the code numbers 052 and 56, respectively.

Our experience is that a SLEX of occupation descriptions without auxiliary information produces too many coding errors. On the other hand, we believe that it is possible to build a powerful SLEX if one can use words of different length and other auxiliary information besides industry.

Sweden is divided into 24 counties and the census coding is carried out one county at a time. When a county has been matched, two lists are made.

The first list is the frequency list, from which an excerpt is given at the top of the following page.

This list contains those occupation descriptions that the dictionary has failed to code and which occur at least twice in the input file.

In that example, it is seen that, for instance, three records with the occupation description "CHARKFÖRESTÄNDARE" (Butchershop manager) have not been coded.

When the coding of a county is terminated, the frequency list is scanned and new occupation descriptions are entered into PLEX. Furthermore, the control lists provide supplementary information for corrections in PLEX. The size of

ADJUNK	052	56	ADJUNKTBIMAKEHÖGSTADIET	93101
ADJUNK	052	56	ADJUNKTBIOLOGIKEMI	93101
ADJUNK	052	56	ADJUNKTBIOLOGIKEMI	93101
ADJUNK	052	56	ADJUNKTBIOMATEMATIK	93101
ADJUNK	052	56	ADJUNKTEKOÄMNEN	93102
ADJUNK	052	56	ADJUNKTENG	93102
ADJUNK	052	56	ADJUNKTENGELSKAFRANSKA	93102
ADJUNK	052	56	ADJUNKTENGELSKAOCHTYSKA	93101
ADJUNK	052	56	ADJUNKTENGELSKATYSKA	93101
ADJUNK	052	56	ADJUNKTENGELSKATYSKA	93102
ADJUNK	052	56	ADJUNKTENGELSKATYSKAMETEMAT	93102
ADJUNK	052	56	ADJUNKTFILOSOFIOCHMATEMAT	93102
ADJUNK	052	56	ADJUNKTFYSIKOMATEMATIK	93102
ADJUNK	052	56	ADJUNKTFÖRETAGSEKONOMI	93102
ADJUNK	052	56	ADJUNKTGRUNDSKOLANSHÖGSTADIU	93101
ADJUNK	052	56	ADJUNKTGYMNASIE	93102
ADJUNK	052	56	ADJUNKTGYMNASIESKOLANKOMVUX	93102
ADJUNK	052	56	ADJUNKTGYMNASIET	93101
ADJUNK	052	56	ADJUNKTHISTORIASVENSKARELIO	93102
ADJUNK	052	56	ADJUNKTHÖGST	93101

## FOLK- OCH BOSTADSRÄKNINGEN 1980

## YRKEN SOM EJ MATCHAT MOT PLEX MED FREKVENSTÖRRE ÄN 1

KOMPRIMERAD YRKESBESKRIVNING	ANTAL
BYGGTRÄ	002
BYRÄDIREKTÖRNATURVÅRD	002
BYRÄINTENDENT	002
BYRÄSEKRFÖRSÄLJNING	002
BÅTTRAFIKÄGARE	002
BÄDDBITRÄDE	002
CEMENTKVARNSOPERATÖR	002
CHARGERARE	002
CHARKARB	002
CHARKARBETARE	003
CHARKFÖRESTÅNDARE	003
CHARKUTERIBITR	003
CHARKUTERISTSTYCKARE	002
CHARKUTERISTTILLVERKNING	002

PLEX increased from about 4,000 records to more than 11,000 during the production.

The other list, the SLEX-list, shows the occupation descriptions which have been coded by SLEX. An excerpt is given below.

In the SLEX-list, coarse coding errors are easily discovered.

SLEX has not increased as much as PLEX, because we have not had enough time to find and try new SLEX words. It contains slightly more than 500 words. As pointed out before, we believe that it would be possible to create a much more powerful SLEX, provided we could use auxiliary information.

The coding degree for the entire production was 71.5%, roughly 68% by PLEX and 3% by SLEX. The coding degree varied between the counties from 67.2% to 76.6%. Our overall goal was 70%, so, everything went slightly better than planned.

The cost for running the matching program is negligible. Look at the example on the following page. The descriptions for one county with 341,529 economically active individuals were matched against a PLEX containing 10,291 descriptions and a SLEX containing 513 words.

The cost for this matching and automated coding was 303 Swedish crowns or about 50 Canadian dollars.

## FOLK- OCH BOSTADSRÄKNINGEN 1980

DATUM 1982-10-06

SIDA 76

## POSTER SOM MATCHAT MOT SLEX

ORDDEL	YRK	SEI	REDYRKE	NÄRG
ÖVERLÄ	031	57	ÖVERLÄKAREKIRKLIN	93310
ÖVERLÄ	031	57	ÖVERLÄKAREKIRURGI	93310
ÖVERLÄ	031	57	ÖVERLÄKAREKLINIKHEFLÅNGVÅRD	93310
ÖVERLÄ	031	57	ÖVERLÄKAREKLINIKERCHEF	93310
ÖVERLÄ	031	57	ÖVERLÄKAREMEDKLIN	93310
ÖVERLÄ	031	57	ÖVERLÄKARERÖNTGEN	93310
ÖVERLÄ	031	57	ÖVERLÄKARERÖNTGEN	93310
ÖVERLÄ	031	57	ÖVERLÄKGYNKOLOGI	93310

	Number of coded records	Coding degree (%)
PLEX	246,652	72.2
SLEX	8,339	2.2
<b>Total</b>	<b>254,991</b>	<b>74.7</b>

It should also be mentioned that, according to our census experiences, the keypunching personnel shall be instructed to punch exactly what is written on the questionnaires (up to a pre-specified number of characters, in this case 30). We believe this gives the best combination of punching rate and quality.

#### 12.4 First manual coding

After the matching against the dictionary, almost 30% of the economically active population remains uncoded. This part must be coded manually. In the census this is carried out in two steps. The first manual coding is carried out on display consoles without access to the questionnaires. Twenty records are shown at the same time on the display console. For each individual, the screen shows the occupation description, the code number for industry, institutional classification, size of establishment, and type of activity.

The coders use an alphabetic occupation list containing more than 12,000 official occupation descriptions with associated occupation and SEI code numbers. The principal rule is that the coder must find "exactly" the same occupation in the occupation list as the one shown on the screen. When the occupation is found in the list, the associate code numbers are keyed on the display. Occupation descriptions which cannot be coded are left uncoded and these records are coded in the second manual coding.

We had predicted that 20% of the total number of records should be coded in the first manual coding. The outcome was 17.1%. The coders involved in the first manual coding managed to code an average of 217 records per hour (including those that were left uncoded).

#### 12.5 Second manual coding

In the second manual coding, the last portions of the records are coded. This coding is carried out on display consoles with access to the questionnaires. We predicted that about 10% of the records would remain at this last stage. The outcome was 11.4%.

The second manual coding is very time-consuming. In fact, this step is very similar to conventional coding of the roughly 10% most difficult descriptions. The coders involved in the second manual coding managed to code an average of 27 records per hour.

As can be seen, this coding in two steps makes the coding life a lot easier. The coding speed of the first step can be kept very high since the coding is carried out without access to the questionnaires.

#### 12.6 Some general remarks

As already mentioned, the resulting coding degree of occupation and SEI in the 1980 Census of Population was 71.5%. Calculations made prior to the decision to use automated coding showed that a coding degree of 60% would be profitable.

It should also be pointed out that the high coding speed obtained in the first manual coding is to a large extent due to the fact that the occupation descriptions are entered into the computer which means that the coding can be carried out without consulting the questionnaires. This saves a lot of time.

Of course, we do not know the exact cost for an imagined system of conventional manual coding of occupation and SEI in this census, but we have reasons to believe that the automated coding saved us at least one million Swedish crowns (approximately 170,000 Canadian dollars), i.e. about 10% of the total coding cost for occupation and SEI.

Money, however, was not the only reason for using automated coding. It would have been impossible to get enough coding personnel at Statistics Sweden to do the coding within reasonable time. Automated coding reduced the number of

records to be manually coded regarding occupation and SEI from about 4,000,000 to 1,200,000 and made it possible to use two manual coding systems.

We also believe that there is a great value having the occupation descriptions entered into the computer per se. As was the case with purchase descriptions, an occupation description contains more information than does a single code number. This "extra" information might be useful to, for instance, future medical researchers.

Unfortunately, no evaluation of the coding quality has been undertaken. However, we know that PLEX results in almost error-free coding of almost 70% of the economically active individuals. The SLEX-lists were carefully checked throughout the entire production process and, bearing in mind that SLEX was used for a few percent of the cases only, we can assume that its relative inaccuracy has no serious impact on the total error rate. We have also, as soon as the coding of one county was terminated, scanned the control lists, i.e. we have listed a sample of records and checked the code numbers in each county. This procedure has led to continuous improvements of the dictionary and the coding instructions. The lists have given us coarse estimates of the error rates and we strongly believe that the occupation error rate is lower in this census compared with the estimated 8% rate obtained in the evaluation of the 1975 Census.

### **13. Other applications**

#### **13.1 Coding of occupation, SEI, and union membership in the Survey of Living Conditions (SLC)**

In the continuing SLC, all numeric and some of the verbal information are keypunched in order to make the editing more efficient. As a by-product, punched verbal information can be used for automated coding. That is the case for the occupation and SEI variables. The punched file is edited and the occupation descriptions are matched against a PLEX dictionary. In case of a match, the code numbers for occupation and SEI are listed together with all the other information punched from the questionnaires.

After that, the code numbers automatically assigned are checked by the coding personnel and changed if necessary. Furthermore, uncoded descriptions are coded on the list. The PLEX used in the SLC is part of the PLEX used in the 1980 Census, namely the part that contains no auxiliary information.

This "semi-automated" system works well and recently it has been extended to include the variable "union membership" (60% coding degree) and there are plans to extend it even further to include the variable "education" as well.

#### **13.2 Coding of occupation in pupil surveys**

Statistics Sweden carries out continuing surveys of different pupil groups. The surveys are conducted a certain time after the pupils have finished their education. The purpose is to get information on their present work and plans for the future.

Almost all the information obtained on the questionnaires in these surveys has always been punched for different purposes. In two recent surveys this punched information has been used for automated coding of occupation.

The PLEX used in the pupil surveys is part of the PLEX used in the 1980 Census, namely the part that contains no auxiliary information.

The coding degree obtained is around 50%. A large portion of the remaining punched occupation descriptions could be manually coded without access to the questionnaires. This coding without consulting the questionnaires is done much faster than conventional coding. This is the same experience that we had with the 1980 Census material; the difference in coding speed between the first and second manual coding is substantial. This is, we believe, an often forgotten advantage of automated coding. That is, the coding speed of the manual part of an automated coding system may be substantially higher than the speed of a conventional, entirely manual coding system.

The experiences of automated coding in the pupil surveys were favorable, and the system for automated coding of occupation is now used in those surveys.

### 13.3 Book loans

The Swedish Author's Fund makes disbursements to authors in proportion to the popularity of their books among borrowers at public libraries in Sweden. This bonus is based on sample data from different libraries and it is distributed once a year. The survey is carried out by Statistics Sweden on a commission basis.

The general data processing situation, where a list of alphabetic keypunched names of authors and book titles is produced, is quite favourable to automated coding. In such a situation it is easy for an automated system to compete with a manual. Even a rather modest coding degree makes the automated system profitable, since the punching is "free of charge". The only requirement is that the computer cost should be less than the manual coding cost on a record-by-record basis.

Only a PLEX dictionary with a 100% unequivocal rate is considered since each error could have a substantial effect on the bonuses distributed.

The system has been used since 1978. During that period the dictionary has increased from 6,900 authors and book titles to 65,000. During the same time, the coding degree has increased from 33% to 80%. Evaluation studies carried out a few years ago revealed that the dictionary containing 65,000 descriptions was not 100% accurate. Therefore, a revised dictionary containing 35,000 descriptions was created. Now the coding degree has dropped from 80% to 68%. Since the system paid off from the

start already, it is now profitable with a broad margin.

### The Future of Automated Coding

Obviously, automated coding is a possible option for some variables when designing a coding operation. We believe that its success is a function of language complexity, though. It seems as if the Swedish language is more forgiving than English in this respect.

In most of our experiments and applications we have used methods that are clearly unsophisticated. Efforts with sophisticated methods have not been especially successful but not especially extensive either. The methodological development has probably suffered from the fact that relatively modest coding degrees around 65-70% have paid off. We should strive for even more profitable systems; we should like the coding degree to jump 10 or 15 percentage points in, for instance, the coding of goods or occupation. This could be done by more sophisticated methods but also by changing the code in some respects. Merging of different categories is sometimes prohibited due to obligations towards the data users. Perhaps it is not too preposterous to make changes in the codes in order to obtain a less costly coding. That option should certainly be considered more often in times of scarce financial resources.

The coding degree can also be improved by storing auxiliary information in the dictionaries and by using more efficient SLEX dictionaries.

Automated coding is here to stay. The Swedish labor market legislation makes it difficult to temporarily hire coding personnel for occasional efforts such as the coding in a census. We have to rely on our permanent staff, and automated coding has emerged as the rescue when it comes to cutting work-load peaks.

So far, our strategy has been to put the easier variables to a test first. Now we have to proceed to the more difficult ones and make the dictionaries and the supporting routines more efficient.

## REFERENCES

- Appel, M.V. and E. Hellerman, 1983. "Census Bureau Experience with Automated Industry and Occupation Coding", American Statistical Association, *Proceedings of the Section on Survey Research Methods*, pp. 32-40.
- Appel, M.V. and T. Scopp, 1985. "Automated Industry and Occupation Coding", Paper presented to the Census Advisory Committee of the American Statistical Association and on Population Statistics at the Joint Advisory Committee Meeting, April 25, 1985, in Rosslyn, Virginia.
- Bailar, B.A. and T. Dalenius, 1969. "Estimating the Response Variance Components of the U.S. Bureau of the Census' Survey Model", *Sankhyā*, Series B, Vol. 31, Parts 3 and 4, pp. 341-360.
- Bäcklund, S., 1978. "Automatisk kodning. Beskrivning av programvara och programvaruhantering", Memo, Statistics Sweden (in Swedish).
- Corbett, J.P., 1972. "Encoding from Free Word Descriptions", Memo, U.S. Bureau of the Census.
- Dalenius, T. and L. Lyberg, n.d. "An Experimental Comparison of Dependent and Independent Verification of Coding", Memo from Tore Dalenius to Leon Pritzker.
- Harvig, H. 1973a. "Kontrollkodning av dödsbevis", Memo, Statistics Sweden (in Swedish).
- Harvig, H. 1973b. "Kontrollkodningsexperiment på blanketter för inskrivningsuppgifter till högre studier", Memo, Statistics Sweden (in Swedish).
- Knaus, R., n.d. "Syntactically Based Classification from Natural Language Responses", Memo, U.S. Bureau of the Census.
- Knaus, R., 1978a. "Inference by Semantic Pattern Matching in Industry Classification", Memo, U.S. Bureau of the Census.
- Knaus, R., 1978b. "Automated Industry Coding - An Artificial Intelligence Approach", Memo, U.S. Bureau of the Census.
- Knaus, R., 1979. "A Similarity Measure on Semantic Network Nodes", Paper presented at the Classification Society Annual Meeting, Gainesville, Florida, 1979.
- Knaus, R., 1983. "Methods and Problems in Coding Natural Language Survey Data", American Statistical Association, *Proceedings of the Section on Survey Research Methods*, pp. 51-60.
- Lakatos, E., 1977a. "Automated I & O Coding", Memo, U.S. Bureau of the Census.
- Lakatos, E., 1977b. "Computerized Coding of Free Verbal Responses", Memo, U.S. Bureau of the Census.
- Lyberg, L., n.d. "Bervende och oberoende kontroll av kodning", Rapport nr 4, Forskningsprojektet FEL I UNDERSÖKNINGAR, Stockholms universitet, Stockholm (in Swedish).
- Lyberg, L., 1981. *Control of the Coding Operation in Statistical Investigations - Some Contributions*, Ph.D. Thesis, Urväl No. 13, Statistics Sweden.
- Lyberg, L., 1983. "The Development of Procedures for Industry and Occupation Coding at Statistics Sweden", *Statistical Review*, pp. 139-156.
- Lyberg, L., P. Nordling, and J. Elmdahl, 1973. "Kodningskvaliteten i läraryregistret", Memo, Statistics Sweden (in Swedish).
- Minton, G., 1969. "Inspection and Correction Error in Data Processing", *Journal of the American Statistical Association*, Vol. 64, pp. 1256-1275.
- Olofsson, A., 1976. "Kvalitetskontroll av näringsgrenskodningen i AKU hösten - 74", Memo, Statistics Sweden (in Swedish).
- O'Regan, R.T., 1972. "Computer-Assigned Codes from Verbal Responses", *Communications from the ACM*, Vol. 15, No. 6, pp. 455-459.
- Owens, B., 1975. "The Corbett Algorithm for Coding from Free Word Descriptions", Memo, U.S. Bureau of the Census.

# THE FUTURE FOR GENERALIZED SOFTWARE

## OR - DOES SOFTWARE GO BAD?

MIKE JEAYS

INFORMATICS SERVICES AND DEVELOPMENT DIVISION  
STATISTICS CANADA

### Background

The census processing consists of four major steps, as listed below:

1. Data Capture
2. Edit and Imputation
3. Tabulation
4. Data Analysis

A sophisticated management and control system is used to ensure the processing is carried out on schedule, and that every questionnaire is properly accounted for.

Generalized software has played an important part in the Censuses of Population conducted by Statistics Canada since 1971. At that time, the STATPAK tabulation system was developed within the Bureau. Its primary feature is a language known as TARELA which can be used to specify a wide variety of tabulations from the census files. (The census files consist primarily of a collection of three relations: one for enumeration areas, one for households, and one for individuals.) STATPAK, especially in its earliest versions, is built around the structure of the census files, and combines a great deal of code that refers specifically to the linkages between these files. It was used in conjunction with a package known as CASPER to produce finished tabulations for publications.

The success of this method prompted serious development of more generalized software for the 1976 Census. The problems of managing a large data base were addressed next. It was quickly realized that no commercial DBMS package on the market could handle files of 24 million records, spanning several disk volumes, and at the same time ensure reasonable retrieval times and efficient use of disk storage. The response was to develop the RAPID system, which uses a transposed file architecture to optimize performance at tabulation time. (Much of the census processing is the production of a large variety of tabulations. The edit and imputation process is also performed on one variable at a time, and the transposed file architecture is of benefit here.) This requirement is significantly different than the needs of most commercial systems, which tend

to be optimized towards record-oriented retrieval, and to provide concurrent update facilities. The transposed file architecture, in which all values of a variable are stored contiguously, is designed to achieve the high level of performance that was required. It has proved highly successful, and RAPID can still outperform all commercial software for this type of processing. No changes in this situation are expected before 1991.

Edit and imputation for qualitative variables (those with a relatively small number of discrete values) was performed by a second generalized package. This is known as CANEDIT, and was developed in time for the 1976 Census. This software was based on a methodology given in a paper by Fellegi and Holt. Edits are specified as "CONFLICT" statements. These consist of a set of logical relationships between variables within a logical record, which are judged to be impossible. An obvious example is that of someone who appears to be married, and 3 years old. The algorithm determines the minimum number of fields to adjust to ensure that it satisfies all the edit conditions, and then fills in the invalid values from a "similar" record which is known already to be error-free. This process is known as imputation. In the 1981 Census, the number of variables collected was increased significantly, from 15 to 46. This made it difficult or impossible to use CANEDIT to perform all the editing and imputation, and a new generalized package, known as SPIDER, was developed. This package accepts edits in the form of decision tables. After performing consistency analyses, it generates code in the PL/I language that is subsequently compiled into a program that will carry out the editing and imputation specified in the original decision tables. While SPIDER is essentially general in nature, a number of extensions were written especially for the census processing, and only later was a partially stripped-down version made available for use on other projects. It should be noted that both CANEDIT and SPIDER were used in the 1981 Census, and will be used again in 1986. SPIDER is capable of doing all that CANEDIT can do, although the form of input specification it accepts is completely different. The two packages together have eliminated the need for writing hundreds of thousands of lines of codes in conventional programming languages,

which would otherwise have been needed. The cost of writing and maintaining these would have been much larger than the costs that were experienced through the use of the software.

An important feature of both packages is that they provide feedback about the input specifications in advance of doing any production work. For example, CANEDIT generates a set of "implied edits" derived from those that were specified. These may be inspected for potential errors; an implied edit that is obviously wrong tells the user that the specified edits must contain an error. These can then be corrected as needed.

The 1986 Census will run, as far as possible, using the same software as the 1981 Census. Very substantial savings have been already achieved by removing the need for redevelopment, and should contribute to smooth running in 1986. This investment has paid off well.

## **Potential Problems and Risks for the 1991 Census**

### **Software Does Not Go Bad!**

Software, unlike many other commodities, does not go bad of itself. However, for a variety of reasons, all software seems to need regular maintenance. The hardware and software environment in which it is used is usually subject to a number of incremental changes. Many of these are usually of low significance, and are believed to be "upward-compatible" – that is, they are thought not to affect existing applications. However, the accumulated effect may sometimes be more serious, with the result that software that has not been used for some time may fail to run altogether when needed. The changes needed to overcome the problems may be minor, but if there are no experienced people available at short notice, the effect on a tight production schedule may be catastrophic. The key to success is continuous use, and thorough verification shortly before a major application.

Besides upgrades to the operating system under which an application or generalized package runs, there will be changes in the computing hardware itself. Current main frames are significantly different to those of 1971. As well as being far more powerful, they have much enhanced instruction sets, virtual memory, more elaborate input-output systems, and much improved random access storage devices.

The operating system's primary function is to insulate applications from most of these changes. The services offered by the operating system have been enhanced over the years, but all the features present in 1971 are still there. This can be reliably predicted to remain true until 1991, and almost certainly for many years thereafter. MVS (or an operating system that provides compatible services) will certainly be around until the end of the century; the investment world-wide in software that runs in its environment is far too large for there to be any real possibility that this will not be the case.

Changes in random access devices pose the greatest need for continual upgrading, and they will certainly continue to increase in performance and capacity for many more years. Read-only devices (optical disks) are likely to be available before long, and they should be highly suitable for a census retrieval system. It may be viewed as certain that any such new devices will be supported under MVS in an upward compatible manner. Market forces will ensure that, provided that software is written in such a manner that it can easily adapt to changes in parameters such as track lengths, difficulties in adapting to new devices will be minimal.

### **Problems of Support for Generalized Software**

Despite all the comments above, it is clearly necessary that software development groups should remain in a position to carry out any upgrades that do become necessary. These can be divided into three categories:

1. Changes enforced by non-upward-compatible modifications to the environment, such as new hardware, operating system changes, and so on. These are clearly essential if the software is to continue to operate.
2. Corrections to problems identified in the software (i.e. bug fixes). New bugs do not appear spontaneously in software – they are either introduced as a side effect of other maintenance, or were there previously, but were not noticed. With the exception of the unusual case where wrong results were produced in the past, but were not noticed, the software can continue to be used for the subset of its capabilities which were appropriate for past runs.

3. Addition of new features. This is, of course, not maintenance in a real sense. While it is often tempting to consider minor enhancements as maintenance, a clear distinction should be made.

For each supported package, the group responsible for general software should keep at least two staff members with detailed knowledge of the internals. While this may not be a full-time occupation, it should be an important part of their work. These people need to be very highly qualified and well motivated. One of the best ways to ensure this is to require them to carry out additional development, often by adding new features to the software, with the firm instruction that whatever they do must be fully upward compatible with previous versions. This expensive activity can best be justified and funded if the software is in continuous use in the organization. Software that is suitable for use in periodic surveys as well as in censuses will be used regularly, and this will ensure that any problems are fixed quickly. Such software will tend to survive within an organization more easily than software that is not used for long intervals. In the latter case, expert staff may have left for work in other areas before the problems are detected, and their absence may increase the time and cost of making the necessary changes.

The development group should be strongly encouraged to ensure that software is written to simplify maintenance. Besides good documentation, this means that there should be an avoidance of "tricks" that are likely to cause problems in the future. Care should be taken to parametrize numbers that are likely to change, such as random access device characteristics. Adherence to these practices in the RAPID system has already ensured its adaptability to an environment that has changed significantly since the earliest version, and this experience should be applied wherever possible.

### **Expertise in the Census User Community**

A similar problem of keeping an adequate pool of trained people within the user community for the census software must be faced. It is difficult to train a new group of people once all the experts have left, as there is no one with the in-depth knowledge of all the details and tricks that are virtually impossible to document. One could do a cost-benefit analysis, weighing the cost of keeping trained and experienced staff with a slight insufficiency of work against the cost of re-training user staff. For packages such as STATPAK, which are in continuous use, the problem does not arise.

But for other software such as CANEDIT and SPIDER, which may not be used regularly, the difficulties are very real.

### **Will there be Better Solutions in the Market-place?**

Commercially available software is evolving at a fast pace, and is becoming more and more sophisticated and powerful. One of the main reasons for the creation of RAPID (that no commercial DBMS could support files of 24 million records) has now disappeared. However, the demands from the commercial world do not correspond closely to the needs of the census. The most significant special demands are as follows:

1. The need for large-scale automated editing and imputation, which is unique to statistical offices.
2. The need for a large variety of special tabulations in a wide variety of formats.
3. The need to produce these tabulations in publishable form of excellent appearance. (This need is being reduced as the emphasis changes from formal publications to the answering of specific requests from the user community.)
4. The need for extreme efficiency in processing files of many millions of records.
5. The need to produce bilingual output.

Some of these requirements are now answered, at least in part, by commercial software. TPL (developed by the U.S. Bureau of Labor Statistics) and the SAS tabulation procedure (itself modelled on TPL) are both suitable in terms of output and flexibility, but do not provide adequate support for bilingual output. So while retrieval and tabulation alternatives may be available, our own edit and imputation software must be developed and maintained.

### **Should We Continue to Upgrade the Software?**

As noted above, one of the best ways of keeping software current is to continually develop and use it. An example is provided in RAPID, which is used by a number of other projects throughout the Bureau, and which has been widely distributed to other installations. A program of continued maintenance and development has resulted in a collection of new utilities, and increased performance.

Skilled developers for generalized software are in very short supply, and can only be trained through years of experience. It is recommended that a particular function be selected for major development, and that resources should be concentrated primarily on software for its support.

For the period 1986-1991, the most important candidate is the edit and imputation function. Requirements for editing have now surpassed CANEDIT. (This was the main reason for designing SPIDER.) SPIDER, in its turn, does not do everything required. It is therefore suggested that a new package, combining the best features of the two, together with any original ideas that have appeared in the last few years, should be designed and written in time for the 1991 Census. This should be done with funds allocated at the Bureau level, with the intention that it should be suitable for use in many other projects besides the census.

The second candidate is the retrieval system. STATPAK is now many years old, and in practice, many tabulations are produced with the SAS system. (STATPAK is used for the initial retrieval from the census files, for reasons of efficiency, and because of its close integration with the geographical selection facilities.) A much more efficient way to interface the SAS tabulation procedures with the census files is a strong possibility for investigation. More generally, the Bureau should concentrate on using commercial software efficiently, by building interfaces or even by contracting with the vendors for extensions that meet our special requirements. This is likely to be more cost-effective than attempting to develop a new tabulation system.

### **New Needs of the Census**

New requirements for the census must of course be studied in detail. New options for data collection will be studied, and the amount of data collected for each person and household is likely to increase further. The basic need for a wide range of tabulations will remain, and it should continue to be possible to generate the majority of these by the use of a generalized package.

New options, such as automatic coding of occupations, should be studied. Recent artificial intelligence methods may be of value in this context, and a pool of expertise should be created within the Bureau.

Alternative means of disseminating information to the public are also suitable for further investigation, especially through the use of modern electronic media.

Cost-effectiveness must also be studied. In general, it may be stated that tuning for performance is much less expensive than rewriting software completely.

### **Review of General Software**

The size of the user community for generalized software varies greatly by package. For example, SAS is by far the most widely used package on our installation, and accounts for about 25% of batch activity. There are very many users, with all levels of knowledge and sophistication. At the other extreme, CANEDIT is used only by the Census of Population. RAPID is used by about 15 different projects at present.

An analysis of usage of each of the packages would make it much easier to judge which should be given more extensive support, and which should be proposed for eventual removal. There are some technical difficulties in measuring the utilization of some of the packages, and solutions should be investigated.

A recent analysis of batch submissions showed that nearly 45% were wholly or partially dependent on one or more general packages. This figure is likely to increase in the future, with the adoption of a proposed policy that strongly encourages their use.

There is a clear need for a review of the collection of general purpose software that is in use in the Bureau. This collection is documented within the software directory, which has been widely distributed throughout the organization. Some packages are very widely used and are well supported; others are used rarely or not at all, and no maintenance has been done recently, except to adapt to environmental changes when these have had an impact that necessitated maintenance. It is proposed that this collection be reviewed in detail by the maintainers and by a group of users, and a report produced to recommend any action needed, such as the formal decommitment of support for unused software. This report should then be circulated widely, to give all potential users an opportunity to comment. It is suggested that the final conclusions, after input from as many users as possible, should be presented jointly to the Informatics Committee, and to the Methods and Standards Committee, for formal approval.

Finally, it is recommended that the funding for general software (including development, evaluation of vendor-supplied packages and support) should be at the Bureau level. It will be driven by the demand from the user community, but not exclusively by any one area. In the past, much of the work done on general software has been

initiated by the demands of the Census of Population. There are pressing needs from other areas, such as business surveys and data analysis. There are likely to be more opportunities for saving development resources if these needs are considered together, rather than independently.



## **Résumé of Discussion**

**Mr. J. Ryten**, Assistant Chief Statistician, Informatics and Methodology Field, provided an overview of the census automation presentations.

### **Introduction**

Automation can be applied to various aspects of the census operating cycle. Processing operations which include preparation of the census, data capture, coding, edit and imputation, tabulations, production of publications and management information reports, seem to be the stage where automation would have the best application. An automated processing operation should improve timeliness of the release of data and reduce staff but may have an effect on costs.

### **Application of Automation**

The application of automation to census processing operations should be based on an assessment of the most appropriate technology for each specific operation within the processing phase, e.g., data

capture operation vs. edit and imputation operation. The determining criteria for selecting particular technology would take into consideration factors related to system reliability and versatility, and associated cost.

### **Role of Past and Future Censuses**

Research and development of automated applications should also be based on both past experiences on technological application and environmental factors in which future censuses will take place. Some caution should be taken when assessing past experiences because of the uniqueness and the infrequency of censuses. New technological environments represent a novelty for each census.

Finally, several issues of concern for future censuses may be increased, for example, public concerns over confidentiality and privacy, concerns over budgetary constraints, like labour and equipment costs, versatility and forms of technology available.



## **Résumé of the Question Period**

*This section summarizes some of the key issues and concerns raised during the presentation on census automation.*

### **Issues Concerning the Implementation of Automation**

*First of all, an automated system has the tendency to standardize data by removing the outliers. This standardization may compromise the data quality of some variables.*

*An automated coding system is particularly useful for variables such as occupation or industry.*

*It is based on a dictionary or library which describes the components of the variable. Sweden and Canada (for the labour force survey) have developed similar automated coding systems. The key differences between the two systems are the code rate applied and the process of updating the dictionary or library.*

*In the United States, an automated system had facilitated the creation of an effective Management Information System (MIS). This system would contribute significantly to improve management control over processing operations.*

*The utilization of both new and old software systems can create problems of reliability and non-compatibility. However, the re-use of a system previously developed is encouraging because of cost factors and resource availability.*

*The automation of processing operations raised two issues concerning human resource involvement. The first issue concerns the increasing need for a specialized and experienced work-force. The maintenance of the existing software and the development of new software require a more specialized work-force than a manual operation. Without increasing the specialized resource allocation, software research and development would be restricted. The second issue concerns the need to reduce the large inexperienced temporary work-force currently used in processing operations. The reduction of this work-force would contribute to cost reductions.*



## **SESSION: COVERAGE AND DATA QUALITY**

Chairperson: Gordon Brackstone  
Director General  
Methodology Branch

Thursday, October 10, 1985



# ISSUES IN COVERAGE MEASUREMENT AND ADJUSTMENT FOR THE UNITED STATES

HOWARD HOGAN

STATISTICAL RESEARCH DIVISION  
U.S. BUREAU OF THE CENSUS

## Introduction

Since the first census in the United States in 1790, there have been problems in finding and accurately counting every person. Improvements in statistical techniques now allow us to identify certain errors in census coverage, including a differential undercount by age, sex, ethnic group and geographic area. Knowledge that a differential undercount exists, and is likely to remain in the future, has led some to propose that the census figures be corrected using statistical information. To address properly the call for adjustment one must address several other issues. On a fundamental level, one must ask: What is a census and what is meant by census results? Other more operational questions arise: How can the estimates of undercount be improved? How can local area estimates of coverage best be made and can they ever be good enough? What are the trade-offs between the most complete field count and the best adjustment? What standards should be used in deciding for or against adjustment? How would adjustment affect public cooperation or the user community? This paper will discuss these issues and the U.S. Census Bureau's research to resolve them.

As measured by the net undercount, census-taking in the United States shows a steady improvement. One series of estimates shows the 1950 undercount was over 4 percent, the 1960 just over 3 percent, the 1970 undercount just under 3 percent, while the 1980 undercount was approximately 1 percent. Thus, over the past four decades, the net undercount has been cut from over 4 percent to 1 percent. Underlying this steady improvement in the national average undercount, however, is a persistent differential undercount. The undercount of black Americans has been approximately 5 percent higher than the national average for every census since World War II. The undercount of black males has been 7 or more percentage points higher than the national undercount for these four censuses. The call for census adjustment arises not so much from a concern for an overall census coverage but a concern for this persistent differential undercount.

Other differentials in coverage exist, although they are not as well documented. Central cities of large metropolitan areas seem to have higher

undercounts, with the undercount falling as both the central city and the metropolitan area become smaller. Rural areas also have high undercounts. Undercounts for other ethnic groups such as Hispanics, American Indians, or Asians seem higher than the national average but not so high as for blacks. There is evidence that the undercount is higher for the poor, for the single, and for the unemployed. Undercount seems higher among those who rent their home than for those who own their home. Undercount is higher for those under the age of five and also for those in the middle years, age twenty through forty-five. It is persistently higher for males than females.

Knowing that an undercount exists and its general direction puts one in a quandary, it is like knowing that your watch is a little fast or a little slow. It does not help very much in knowing how to set it. Our users previously accepted the census numbers as absolute truth or at least as the best figures available. Now they and we are confronted with fundamental questions about just what a census is and what it would take to make it better. The dilemma is best illustrated by the United States Constitution which in one place calls for seats in the House of Representatives to be allocated based on the whole number of people, and in another place calls for an actual enumeration. In the framers' mind there was no conflict between these two principles, because no alternative to an actual enumeration existed. Alternatives in the form of statistical estimation now exist, and we are now confronted for the first time with this conflict.

Barbara Bailer has written of the two models of census-taking, the Participative Model and the Statistical Model. In the Participative Model, accurate results are achieved through the participation of newspapers, magazines, television, radio and community leaders. The Participative Model holds that participation in and of itself is important. This is the census as a national ceremony. Clearly, the majority of people who fill out a census form, and indeed, the majority of users have this model in mind. And it may be that in the long run the good of the republic is best preserved through wide participation in the census.

However, inaccurate census results hurt everyone, not just those who chose not to

participate. The Statistical Model of census-taking holds that the good of the republic is best preserved by having the most accurate census estimates possible, even if these estimates require sometimes arbitrary and unverifiable assumptions. Rather than using resources to attempt to get all people to participate, the Statistical Model directs some resources towards efficient estimation and the appropriate adjustment of census counts.

If we may talk of two models of census-taking we could also talk about two theories or two views of census results. One view holds that it is the duty of the census taker to report the data as collected. The other view would report population estimates whether they are directly derived from the census or not. This latter viewpoint would widen the discussion of adjustment far beyond the adjustment of census counts, the adjustment of the census results, or even the characteristics of missed people. It would attempt to adjust the reported characteristics for known inaccuracy. For example, if building records clearly showed that the majority of structures in a block were built before 1940, what purpose does it serve to publish the misreports of the current residents as to the age of the buildings? The best estimates would be based upon the best data.

Up to World War II, virtually every census was based on the participation model and the results were clearly census data. Since World War II, there has been a drift, at least in the industrial countries, away from actual responses toward estimates. Hot-deck imputations, substitution, allocations, mover corrections, occupancy status corrections; all are clearly drawn from a statistical model of census-taking. However, our movement has been very timid and we have not strayed far. The calls for statistical adjustment for undercount seem not to be of the nature of marginal adjustment, but rather of quantum leaps into a new viewpoint and a new census-taking methodology. Do we know enough to correct the census? Have we progressed to the point where results based on statistical estimation are clearly superior to those based on the maximum amount of participation?

### **The Best Enumeration or the Best Estimation**

Assume for the moment that a census planner's only goal is to achieve the best possible population estimates for small areas. How many resources would be put into the field enumeration, and how many into the coverage measurement study? First, what do I mean by resources?

Clearly, money is an important resource. Money can be taken from the field enumeration and put

into the coverage measurement and adjustment program. However, beyond a certain minimum, money is seldom a real constraint for the coverage measurement study; neither are its requirements large in relation to the costs of the basic enumeration. In 1980, the budget of the Post Enumeration Program could have been doubled with no major improvement in accuracy. The difficulty of maintaining the high quality of interviewing and matching of a coverage measurement survey places very real constraints on the sample size. These constraints are reached long before the total costs become significant relative to the cost of the basic enumeration.

What other constraints are there? The most important constraint is time. Time here has two different effects. First it takes time to conduct the field enumeration. The more time the census takes, the more people can be counted, other things being equal. But conducting the coverage measurement study also takes time to do the interviewing, matching, tracing and estimation. If it is important to have the census results within, say, nine months after Census Day, then there is only so much time to go around. The planner must decide how to divide this time to produce the best results. Another dimension of the time resource is proximity to the reference date. The closer the census is conducted to the reference date, the better the results. Similarly, the closer the coverage measurement study is conducted to the reference date, the better the responses and the better the quality.

The other resource is staffing, especially of trained field and clerical staff. If the planner were willing to conduct the census enumeration and the coverage measurement study at the same time, then it would be necessary to recruit two temporary field staffs and two temporary clerical staffs. In the United States, we have difficulty recruiting even one temporary staff at a time.

The United States and Australia give two extreme solutions to this problem. In 1980, the United States attempted to solve the undercount problem by concentrating upon the basic enumeration. Special procedures were added. Field offices remained open five or more months after Census Day. The effect was a marked decrease in the average undercount, including a decrease in the average undercount for blacks. Australia, by contrast, finished its field work in **two weeks!** A post-enumeration survey is carried out almost immediately, and then the census is adjusted. Of course, the United States contains areas which are far harder to enumerate than anything that exists in Australia, so no real comparisons are possible. However, it is certainly worth pondering which approach is more likely to result in the

better population estimates, and also, whether there might not be an intermediate position which would be optimal. We also need to think about what testing and research would lend as an empirical answer to these questions.

For 1990, the U.S. is pursuing an intermediate position. The United States has adopted a policy of a complete count and concurrent evaluation of the completeness of the count. We want to have the best field counts possible and the best adjustment possible by our legally mandated date, December 31, 1990. We will then compare the two data sets and decide whether the adjustment meets our standards. This policy is adopted because we have no way of knowing or reasonably predicting whether the adjustment will meet our standards or even be completed on time. In 1986, we plan our first test of this two-track approach. We plan to conduct a full-scale census test with concurrent evaluation. We also plan to prepare adjusted tabulations and publications. Obviously, this is not a test of adjustment accuracy, since one site alone is insufficient to measure this dimension. It will, however, allow us to test the feasibility of adjustment-related operations.

#### **How Can Estimates Best Be Made?**

This is not the place to review all the available techniques, their strengths and weaknesses. Since 1980, we have given serious attention to each of the major techniques: post-enumeration surveys, reverse record checks, administrative record matches and demographic analysis. I will, instead, make a few general points.

First, which measurement technique is the best depends upon the nature of the country, the nature of the government, the nature of the census and the nature of the undercount. The size of the population, the extent of mobility and the extent to which people know their neighbors will affect the success of both the reverse record check and the post-enumeration survey. In a nation such as the United States, with a decentralized government, record systems tend to be scattered. State and local privacy legislation prevents or delays access to important record sets. This obviously makes administrative record checks more difficult. In the United States, birth records are kept by the states, rather than the federal government, so it is difficult to draw the birth sample for the reverse record check.

One must not overlook the nature of the census and the nature of the undercount in deciding what is best. Do omissions constitute the main coverage error, or is net error made up of some omissions and some overcounting? How close to

the reference date can the coverage measurement study be conducted? Are omissions due largely to errors in carrying out census procedures, or are they also due to conscious avoidance of the census? Conscious avoidance of the census will lead to bias in any method which requires talking directly to the person. I should also point out that the quality for the reverse record check depends upon the quality of the previous census and the success of the previous coverage measurement study.

Next, which measurement technique is the best depends upon the problem at hand. Experience in the United States has shown that demographic analysis is often the best approach to measure undercount nationally. Further, as one might expect, demographic analysis is usually best to measure undercount for demographic groups. Almost all that we know about undercount for adult black males or young black children derives from demographic analysis.

Reverse record checks have great advantages for deriving undercount estimates for large geographic areas. Large, of course, means large relative to the internal migration rate. For smaller areas, census blocks for example, it becomes difficult to draw a sample to represent the current population rather than the population five or ten years ago. The post-enumeration survey (PES) is probably the best to understand what is happening in an average block during a census. One can cover all Census Day residents. One can see who was enumerated, who was missed, and who was counted in error. The PES may still be the best for measuring overall census error, at low cost. However, by itself, the PES does not succeed in measuring some types of systematic error, for example, the omission of single black males.

Administrative records may still be the best for targeting certain groups. Thus, if the purpose of the coverage measurement is to evaluate quality of the census and nature of the undercount, administrative records must be considered. For any particular group, drivers, veterans, taxpayers or whatever, there is a set of administrative records which could be used for evaluating the census. The problem, perhaps the insurmountable problem, for administrative records, is how one can integrate the separate parts to form a probability sample representing the total population, especially the total population for a small area. Erikson and Kadane believe that they have solved this problem through their approach to building "composite lists". However, I certainly believe that the problem is not yet solved.

Our undercount problem in the United States is such that no single method seems completely

right. We clearly want direct estimates for a sample of small areas; we will need these in modeling. However, we also have a very severe systematic undercount, as mentioned above. We are working to perfect the post-enumeration survey, because, for 1990 at least, we believe that some variation of a household survey is the most workable approach on a large scale for very small areas. We are also working to see if the PES can be used in combination with other methods, such as demographic analysis or administrative record checks, to improve estimates for hard to enumerate groups.

Let me describe here some of our research into coverage measurement techniques. Rather than being comprehensive, I will just mention a few of the most interesting.

The Forward Trace Study is an experiment aimed at testing the feasibility of the reverse record check technique work in the United States. It has been going on since 1980 and just finished in the field this month. In 1960, we found that waiting to start the tracing until after the next census resulted in too high a non-trace rate. With a large country, a highly mobile population and a ten-year census period, the trail was just too cold. So, we have started an experiment into forward tracing. Instead of waiting until the end of the period, we started our tracing right away.

The Forward Trace Study is actually an experiment with three treatments. In the control treatment we use the techniques used in Canada, waiting until the end to gather tracing information. In another treatment, we interviewed at the beginning to gather information on contact persons, that is relatives and friends, and also social security number to allow access to administrative records. In the third treatment, people were interviewed periodically during the tracing period to help keep track of them. In both of the active tracing groups, we checked administrative records and post-office records during the period to update addresses.

The study includes all parts of the reverse record check sample: a sample of census enumerated people, births, immigrants and people missed in the last census. The sample is heavily weighted toward groups for which the PES normally does badly. Without a mid-decade census we cannot match to anything, but we can evaluate our ability to maintain people in a sample over time. The final results are not in, but it does seem that active forward tracing does yield significant results, and should be considered by any nation using the reverse record check approach. What we still do not know is whether the tracing is successful.

We are experimenting in using administrative records to improve coverage in a post-enumeration survey. Since 1960, census-taking techniques have greatly improved through a variety of coverage improvement techniques surveys. However, our thinking on post-enumeration is stuck in the era of conventional enumeration. We want to use administrative records to improve the coverage of the **post-enumeration survey**. In this way we hope to include some of the systematically missed population.

So far I have only mentioned **post-enumeration** surveys. One also could conduct a **pre-enumeration** survey. In this way, some of the critical timing issues can be finessed. On the other hand, there are real concerns about sensitizing the population as well as the census field staff with this technique. We are planning a small experiment with the pre-enumeration survey for 1986.

Finally, the most exciting is our work on automated matching and computer-assisted clerical matching. The automation of the 1990 census gives us opportunities which never before existed to match on the computer. The savings in terms of time and money are obvious. There are also important gains in improving matching accuracy and reducing matching bias. Using the theory developed by Fellegi and Sunter, our matching research staff has developed a surprisingly powerful computer matching program. We will test this program for the first time with the results of the 1985 Test PES. For 1986, our goal is to integrate this system with an on-line computer-assisted clerical matching system, so that probable matches and other problem cases can be resolved rapidly.

### Small Area Estimation

A census is important, mainly because it provides data for very small areas. Many important uses of census data require estimates for very small areas: legislative redistricting, fund allocation, school enrollment projection and highway planning, to name a few. For states or provinces, other cheaper estimates are statistically, if not constitutionally, as good.

Here then is the dilemma. Our methods of measuring the undercount work best for larger areas: states, big cities, or the nation. Thus, for the large geographic area, we may know an undercount exists. Even for a given stratum of similar areas, we may know that a large undercount exists. However, the best population estimates we have for each individual block may be the census field count.

Actually, this practical, if not statistical, paradox is most acutely posed when one thinks of overcount. It is quite possible that based upon the best evidence and models, one would believe that a certain class of areas was overcounted. Still, the best population figure for each and every block in that area might be the population counted there.

Our current research in this field centers on the use of synthetic estimators integrated with a sampling design to maximize their efficiency with regard to the undercount. Rather than concentrating upon direct estimates for states and large cities, we are considering aggregating across states to form strata which are homogeneous in terms of characteristics we believe to be related to undercount. These would be things such as housing characteristics, racial and ethnic composition, income and education. We are looking at the results of the 1980 census to investigate this approach. Since undercount estimates are not available for small areas, we are first using census substitutions, that is whole person imputation, as a surrogate variable. We use substitutions since they not only are correlated with undercount, but they have the same magnitude. We can thus test the predicted rate using this approach against the known census result related operations.

### Adjustment Standards

Implicit in this decision is an assumption that one knows when the adjusted data are superior to the field counts. However, what does one mean by "better"? To answer this question, one needs a loss function by which to compare the two alternative data sets. Assume that one has a data set which is the result of the best field enumeration possible. Assume also, that one also has another data set which is the result of the best adjustment possible. Assume finally, somewhat implausibly, that the truth is known. It is unlikely that one or the other data set will be closer to the truth for all areas and groups. Thus, one needs a loss function to weight these errors, even in theory.

Mean squared error, mean absolute error and low maximum error are all examples of loss functions. Ideally, the chosen function would be based upon the actual uses of the data. That is, the shape of the loss function would reflect society's real loss due to census errors. This is not an easy task, even for just one particular census use. To sum meaningfully over all principal uses is a challenge.

Of course, adjustment standards cannot be based upon the premise that the truth is known. Decisions must be based upon observable data. Such data might include the measured coverage

errors, the level of differential undercount or the non-response rates in the coverage measurement survey. These observables imply error in the two data sets and therefore imply given levels of the loss function. An important part of the research, then, is determining how to relate the observables to the loss functions. This means, first, constructing error profiles for the coverage measurement estimates, as well as for the census. It means finding methods to measure the errors, or at least assess their magnitudes. Finally, it means finding a way to relate these measured errors to the loss functions, either analytically or through simulations. The United States Census Bureau is starting a major research effort in the area. We are attempting to enlist statisticians throughout North America to help us. The challenge is great, but any progress which can be made will get us far closer to understanding the nature of the decision.

Of course, as I mentioned above, the census is not a purely statistical exercise. How would adjustment affect public cooperation? There is very little experience, but there is some available evidence. The Australians, I am informed, have noticed no deterioration in cooperation after they decided to adjust. In the United States, the public seems unconcerned that we impute people for houses when we cannot determine the number of residents. In 1970, the United States imputed many people into the census based upon a statistical study of people missed in units mistakenly enumerated as vacant. The public was unperturbed. I believe that what is important is that the adjustment be seen as a natural part of the census. If it is seen as substituting for the census, changing the census, or otherwise lessening the importance of the census, then it can cause real damage.

Similarly, the results must be presented in such a way as to make it easiest on the census user. Assuming, as we must, that any adjustment which would be made would improve data quality, this should please the user. Ignorance may be bliss, but it can also lead to costly decisions. Users like tables that add up and results that are consistent. To achieve consistent tables the adjustment must be carried down to the lowest level; indeed the simplest way may be to impute the missing people into the census micro records. This presents a problem. We know how to measure population undercount (at least in theory) and how to impute for it. We know how to measure housing unit undercount and impute for that. We know how to measure whole household undercount, although this is harder, and impute for that. What we do not know is how to do all of this simultaneously and keep everything consistent. How, for example, does one handle within-household misses?

I am sure that these problems can and will be solved, although perhaps not by 1990. Lacking a complete solution, ways can be found to present the data in a meaningful and useful manner. However, adjustment would represent a change for the user. A census organization is obligated to institute a program to educate the users about the changes.

### **Conclusion**

The United States Bureau of the Census has not yet developed adjustment standards, much less

decided that an adjustment can meet those standards. The Bureau is committed, however, to a vigorous program of research to set those standards, and, if at all possible, to meet those standards. This research program has already gained greatly by cooperation with Statistics Canada. We certainly hope that our research and our insights can help in planning the 1991 Census of Canada. We certainly have already benefitted from your experience.

# MAKING DATA QUALITY ASSESSMENT MORE RELEVANT

RICHARD BURGESS

CENSUS AND HOUSEHOLD SURVEY METHODS DIVISION  
STATISTICS CANADA

## Introduction

The subject of data quality assessment and its role in the census have been and continue to be a subject of considerable discussion and work. Approximately 5% of the cost of both the 1981 and 1986 Censuses is allocated directly to data quality assessment activities. Despite this, however, it is not clear that the cost or the specific activities will continue to be justified. This paper addresses the issue of the relevance of data quality assessment in the Canadian Census.

The next part of this paper briefly describes the environment in which data quality measurement and assessment activities must be planned and carried out. The subsequent parts describe the current status of activities within this environment and discuss what it is that various users require of a data quality program followed by a few suggestions on how the relevance of data quality assessment can be improved.

## The Census Environment

Data quality measurement for the census is carried out in a unique environment. In recent years, one of the primary characteristics of this environment is that of limited resources. Resources are limited not only in terms of dollars or person-years available, but also in terms of knowledgeable and experienced staff.

Second, the census has evolved over the years into a production-oriented process. Problems or changes are perceived first and foremost as a threat to the schedule and/or budget. The overwhelming majority of census staff (for example, there are about 40,000 enumerators) are concerned with a given set of procedures and the application of these procedures within a given time period. These people must make on-the-spot decisions, and although their intentions may be the best it is frequently the case that they have no clear understanding of the impact of their decisions on the quality of the final product.

Next, it must be recognized that the data collected in the census must have some utility. In some cases, this requirement is at odds with the desirability of high quality data. A good example of this conflict is the structural type variable collected in 1981. The detailed breakdown of

codes produced very poor quality data. On the other hand, collapsing of the codes, while it resulted in much better quality data, makes the data much less useful.

Finally, it must be kept in mind during the planning, implementation and evaluation that there is always the underlying fact that Statistics Canada will ultimately be held accountable for the quality and the usefulness of the census data.

## The Current State of Data Quality Measurement

Within this environment we may now examine the current state of data quality assessment.

In terms of the technical problems, one of the other papers in this session (by Howard Hogan) gives some excellent examples of some of the problems associated with data quality measurement. These types of problems extend beyond the measurement of coverage error, although in the census context over 75% of our expenditure goes for coverage evaluation.

It is internationally accepted that data quality measurement and the dissemination of information about the quality of data should be provided. Within Statistics Canada, this is recognized by official policy. In the census, the analysis of data quality is carried out both through direct studies and through the certification process. The two of these together are designed to provide a basic understanding of what the data actually represent.

At the same time, however, data quality activities are sometimes perceived as a threat to the schedule. There is a danger that release may be delayed if a problem is identified and a necessary cautionary note is not completed as quickly as dictated by the schedule. There is also competition for the census forms. At the same time the data are being captured, the Reverse Record Check, the Vacancy Check and other evaluation studies all want access to the census documents. In this context, data quality measurement is sometimes not viewed as being a real part of the census. It only becomes so when something goes wrong or someone criticizes the data.

Turning to the question of how data quality measures are actually used, it is clear that there

are some real problems. For example, the formal incorporation of data quality measures into a decision-making process often can lead to severe complexities. Because of this, there is a tendency to ignore data quality measures, even when the measure may be as basic as a sampling error. It is far easier to take the data as is and do a straightforward analysis. It is a common perception among producers of data quality measures that users are often not interested in data quality. They are more concerned with just having some data. Even if they were interested, they are often not technically able to use whatever measures we produce.

In this regard, it has to be acknowledged that we provide little assistance to users on how to actually use data quality measures. For example, we produce extensive amounts of information on coverage error, but we ourselves do not make any general adjustments for coverage error. What then do we expect the users to do with the measures we give them? All we do is encourage, in general terms, the use of measures of data quality. In some cases, such as with response rates, this may be all that is possible. It may also be consistent with what the user wants, in the sense that some users only want to know if the data are useable or not and nothing more. However, in most cases, we leave it to the user to decide on his/her own what to do with the data quality measure.

Finally, it should be noted that data quality studies tend to be very demanding of the resources which are applied to them, in particular the staff who work on these studies. They have to balance the demands of producing the data quality measures with the need for thorough analysis. Production of data quality measures cannot be viewed simply as a production process, since the demands for accuracy in the results are particularly great. There must be very clear and supportable documentation for what is said by these studies.

### **User Needs for Data Quality Information**

There are a number of users of information on data quality in the census, each with their own specific needs. The remarks in this section are somewhat speculative, since in fact it is often not clear what information users want.

For data users, that is those who use the data which are produced as output from the census, they want most of all some easily understood statement or measure of the quality of the data. They particularly want information on the quality of data for small geographic areas and they need

some assistance in determining how to use these particular measures or statements.

A second group of users are census managers and planners. Senior management and planners may have to consider questions such as should the postal check be conducted and where, where should self-enumeration be used, or where should early enumeration be conducted. The answers to such questions depend in part on the associated data quality. For the most part, however, clear answers cannot be given because we do not have adequate information on what the impact on quality would be.

Planners are particularly interested in building quality into the census. Once the census starts to move, nothing is going to stop it. Therefore, any quality that we have is initially provided by the respondent, which demands that quality be built in through such things as a good questionnaire and good collection procedures. Quality is also very dependent on what a large and mostly temporary staff then do with those data. Therefore, we must not only develop good procedures but have good methods for controlling their work.

A third group of users are the individual staff members within Statistics Canada. Those doing analysis and specifying content for the next census must have a good understanding of the data, including their quality. Operational staff should also understand the significance of their involvement in terms of its impact on data quality, since this will lead to better motivation. For example, the Regional Office staff use coverage error estimates as a way of indicating whether they did a good job in a particular census. If a person in Regional Office Processing or Head Office Processing can understand that their decisions can have a direct impact on data quality, then they may be motivated to think more carefully or to do more consultation before coming to a final decision.

Lastly, Statistics Canada as a whole benefits from data quality measurement, in the sense that it supports the integrity of the organization. It is always very nice when outside users criticize the data to at least be able to say that we have examined the quality of the data, even if we cannot support them as being of high quality.

### **Improving the Relevance of Quality Assessment**

Because we do spend considerable resources on the measurement of data quality, we must ask ourselves if there is anything more we can do to improve its relevance.

First, more information is required on what it is the data users need, what they can understand and what kind of analyses are actually done with census data. It is very difficult to provide the user with useful measures if we do not know what he wants to do with them. For the 1991 Census, this could be added as a feature of the consultation process for the output program.

We can also make improvements to the methods of presentation of data quality information. Simply by presenting the information slightly differently, for example in graphical form, we could get the user's attention much more effectively.

In terms of the forums in which we disseminate data quality information, we could consider alternatives to our present approach of producing a few large and complex publications. We should instead have a kind of summary guide of no more than 20 pages in which we present in very brief form the methodology of the census as it relates to the quality of the data, and then go variable by variable into giving a very brief indication of how we feel about the data, i.e. is it very useable, of questionable value, etc. In some cases we may refer users to other sources or caution them that they should obtain more information on these particular variables. Such an approach would get to more people and have a greater impact than the sort of thing we have done in the past.

We should also provide some indication of the variation in quality from one geographic area to another. Typically, what we have done in the past is present a series of results followed by the statement: "the actual rates may vary from one geographic area to another". This is no doubt true but it is not very useful. It may even be viewed as some sort of escape clause on our part.

We should also make some attempt to integrate data quality measures into our own analysis activities. It can be argued that we do not do much analysis or that it would be more expensive. However, we could do some type of summary or

compendium which would show users how they could use some of this information in their data analysis.

Finally, we should re-emphasize the role of data quality measurement in influencing the shape of the next census. This is a role which has become very difficult to defend in recent censuses because of the increasing squeeze on resources. It is very difficult, for example, to convince the 1986 Census manager to sponsor something that will only show results in 1991. However, in the past this has been one of the areas of biggest pay-off for data quality evaluation. Data quality measures were involved in such decisions as going to self-enumeration, in deciding on 1 in 3 and 1 in 5 samples, and to a large extent to the development of such things as the CANEDIT and SPIDER systems. It is essential that we direct ourselves to that type of activity, because it appears that we have reached a plateau in terms of data quality.

In concrete terms, there are a number of ways in which we could do this. We can go out and determine the real cause of errors. This is particularly important in terms of coverage, where what we really have are the symptoms and not the real cause. This information could not only be used to improve the next census but would also be of value in the adjustment issue. We could also expand the particular data quality studies we do, especially the coverage studies. We could collect several other things at the same time; for example, we could ask people what they really understand by a mother tongue or ethnic origin question and whether they can respond in a manner which is consistent with the response categories we have given them.

In summary, we must re-examine the data quality program in terms of its priority and directions. The program does represent a sizeable investment, but with a small amount of change and redirection the measures which we do produce can be of considerably more value to both outside users and for inside planners.



# ADJUSTMENT FOR NON-COVERAGE ERRORS

CHRIS HILL

CENSUS AND HOUSEHOLD SURVEY METHODS DIVISION  
STATISTICS CANADA

## Introduction

Procedures used by the Canadian census for the adjustment of missing and inconsistent data are among the most sophisticated and also the most extensive in the world. In this paper, the rationale for this practice is considered. The case for and against this being done by Statistics Canada is described, and modifications which should be considered are outlined. Three alternatives for 1991 are proposed. The paper will not describe specific imputation methodologies or systems used in the census, as these are well documented elsewhere.

## The Need for Adjustment

Data must be adjusted for missing and inconsistent responses. This step is an integral part of the survey process. It is necessary to move from the raw data to some final data which are considered in some sense to be official or at least approximately true estimates. Various methods are available to do this: edit and imputation, that is, adjustment of records at the micro level, reweighting of data, or even ignoring the erroneous records. It is important to recognize that ignoring erroneous records is in fact a data adjustment process.

Ideally, what one attempts to do is to obtain true estimates. In practice, what is achieved is a data base that is clean of the inconveniences of missing or inconsistent data. Why can we not do better than we actually do? The answer is that a lot of errors are in fact undetectable in the micro data. They can only be detected by subsequent evaluation. Some errors may introduce bias into the data while others may not, but unfortunately at the time of producing the census data it is not known what those biases are. By the time the evaluation is available, it is too late to make adjustments. Because we adjust for some things but not others, a case can be made for moderation in what we do. It is fair to question why we put excessive resources and energy into doing something that is only a partial solution to the problem.

By way of illustration, consider data on mother tongue. In 1971 and 1981, we published figures for mother tongue which were adjusted for non-response. In 1976, we left a not-stated category.

As a result, at the time of publication of the 1976 data, both the English and the French press reported that for the first time in history the percentage of people of francophone mother tongue had declined below 80%. However, if we make even a rough adjustment for the not-stated category that was published in 1976, we find that, in fact, there is a steady increase of the percentage with French mother tongue in Quebec.

What is at issue is not whether data should be adjusted. Adjustments have to be made at least for missing data, if not for inconsistencies. The real issue is whether or not Statistics Canada should be doing it or whether it should be done by someone else.

## The Case Against Statistics Canada Adjusting

First, let us consider the case against Statistics Canada doing the adjustment.

One reason is financial. Adjusting data is expensive. In 1981, it was approximately 10% of the total cost of the census.

An even stronger argument against Statistics Canada adjusting is the timeliness of the process. Approximately four months were added to the process for the 100% data in 1981. For the 20% sample data, approximately a year was added. From this perspective, the few short weeks it takes to collect the data is trivial.

A third point to consider is the riskiness of the process. Both the 1971 and 1976 Censuses had very serious delays in release of the data because of problems in edit and imputation. In 1981, it was a very near thing several times.

The fourth point is that we have in fact been accused of fudging the data. In various academic articles and in various conferences I have attended, the concern has been expressed that we are not being truthful, that we are instead giving data which have been massaged to make them look better. There are two parts to the fudging argument. One is that we are really distorting the data in some way that the user is not able to follow. The second is that some users understand what we have done but disagree with the specific adjustment we have made. A good example of this

is the language data, where some of the adjustments we have made have been strongly criticized.

### **The Case for Statistics Canada Adjusting**

Strong arguments can also be made in favour of our continuing practice of adjustment. The first of these is quality. Statistics Canada can do a better job of the necessary adjustments because we alone have access to the micro data and can take all of this information into account when making the adjustment. Related to this is the fact that we have by experience built up a considerable body of expertise on how best to do it.

A second argument in favour is that of consistency, to which there are three aspects. One is consistency within and between tables for a particular census. If these adjustments are not made, there are inconsistencies between tables that cause users a great deal of inconvenience. Second, there is the question of consistency across censuses. This is an argument not only for our doing it, but for our doing it in a consistent manner. The mother tongue example presented earlier clearly shows the pitfalls of inconsistency. The third aspect is the fact that census data have official importance in terms of various legislative and administrative acts, and that the census data that are published must be consistent for all of the various parties to these administrative and legal arrangements.

A third argument in favour of Statistics Canada doing it is to view it as a service to users. Many users of data do not wish to deal themselves with the problems and complexity of incomplete data. It is also probably more efficient for us to do it once than for many users to do it individually. Even within the organization, for the persons disseminating the data to have to wrestle with the problem of inconsistencies in the data they are trying to produce is a major problem.

The final argument in favour is one of credibility. It is sometimes argued that oddities in the data can undermine confidence in the data themselves. My personal opinion is that this is a poor argument.

### **Three Options for 1991**

In light of these arguments for and against Statistics Canada adjusting data, three options may be considered for 1991. The first is that we continue our traditional practices but that we focus much more attention on what may be called "risk management". The second option is that we consider the preliminary release of some unedited

data. The third option is that we limit the amount of adjustment that we do, that is, that we in fact exclude some variables from adjustment that are currently adjusted.

In terms of our traditional practices, a brief bit of history is informative. Since the 1971 Census, Statistics Canada has gone very much for the practice of automated edit and imputation. In 1971, there were a lot of problems, and there was an attempt to rectify these problems with the introduction of the CANEDIT system in 1976. The CANEDIT system had two important characteristics. It had a very rigorous methodology underlying it, and it placed a very strong control on the development and implementation processes. However, when we came to 1981, we recognized that in some areas CANEDIT placed unnecessary constraints in terms of the flexibility of adjustments one could do. We also recognized that there were certain variables, basically those with long code lists and arithmetic variables, which CANEDIT could not handle. Because of this, the SPIDER system was developed. These four points, namely the rigour, the control, the flexibility and the range of variables are in my opinion the four things that one needs in an edit and imputation system. The question is whether you can get them all in one package.

Dealing specifically with risk management, there are four recommendations. First, we should re-use CANEDIT in 1991. Second, we should make SPIDER more rigorous. Third, some of the complexities of the edit and imputation strategies need to be rolled back. Finally, more work is needed on how we control the whole process. Although we made tremendous progress in 1981 in controlling the whole imputation process, we need to work on harmonizing that component.

Turning now to the more controversial issue of preliminary release of some data, it is clear that this would be a break with tradition in the census. There are one or two small precedents, such as the early release of special tabulations for the Yukon and Northwest Territories, but these were essentially unofficial. However, there are other countries where unedited data are released. Israel, for example, releases unedited data and subsequently releases edited data. Australia also does a much more modest form of edit and imputation than we do and releases data that we would term as unedited.

Another interesting point to note is that it is common practice for the economic side of Statistics Canada to release preliminary figures followed by adjusted figures, while on the social

side the general practice is not to do so. We delay our data a lot longer and then produce final data. The principle behind this preliminary release of data is worth examining in the census. Twelve months is a long time for the data to be sitting on the data base, especially when some users know they are sitting there.

Of course, a valid concern in this issue is cost. We would have to examine the additional costs of this preliminary release and the logistical problems that they caused. However, it should be recognized that with the current process we do produce a large volume of monitoring tabulations for the purpose of deciding whether we are doing the adjustment process properly. A lot of these monitoring tabulations might in fact be suitable for release. What is suggested is not a glossy or expensive publication but rather a preliminary, cheaply produced release that is disseminated with the caveat that it is a quick tabulation of the data which may have problems within it.

A very real concern with this approach is credibility. Concern has been expressed in the past that if people see the data with all of the oddities in them, we will lose credibility. However, I feel that the reverse argument is true. When people see the unedited data as well as the edited data, they will find that what we do has a very modest impact on the data and that what we are really doing when we adjust the data is to make them more convenient and easy to use rather than making any major modifications to them.

The third option, that of cutting back on the amount of adjustment that we do, is in a sense related to the second option. However, instead of

releasing preliminary data and subsequently going on to edit them we could seriously consider not doing any further editing or else making the adjustment with a very simple procedure such as weighting up for non-response, which is cheap and easy to do. There are obviously benefits of cost and timeliness with such an approach.

In terms of how users would feel about it, it is helpful to divide the user community into two groups. The first group is those who just want simple counts, often for a small area, without any complicated explanations of how we processed the data. On the other hand there are a few academics who are interested in doing some very complicated analyses of the data. The detailed adjustments we do now satisfy neither user. For the user who just wants simple counts, a complicated explanation is wasted; a very simple form of adjustment would meet their needs just as well. The sophisticated user is unsatisfied as well, since he may disagree with our approach and would prefer to have the unedited data to do his own analysis. We have a situation, for the cultural characteristics at least, where what we now do does not satisfy our users very well.

### Summary

In summary, three recommendations can be made which address the issue for 1991. First, we should concentrate on risk management, in particular not introducing anything into the adjustment process that we don't need to. Second, we should consider preliminary release of at least some of the key figures. Finally, we should seriously look at areas where the adjustments which we do should be limited.



# ADDRESS REGISTERS: ADVANTAGES AND DISADVANTAGES

DAVID C. WHITFORD

DECENNIAL PLANNING DIVISION  
U.S. BUREAU OF THE CENSUS

## Introduction

For the 1970 Decennial Census of Population and Housing, the U.S. Census Bureau adopted for the first time a mail-out/mail-back enumeration method for most large urban areas. In these areas, a list of all living quarters was compiled and census forms mailed to addresses on the list. Addresses on the list were printed in address registers which were used to control the enumeration process. Each address register (AR) contained information for living quarters for a geographic area (about 300 units). During the census process, ARs were used to indicate corrections, additions and deletions of addresses, as well as identification of householder name and physical location description, receipt of questionnaires, population counts, and content sample designation.

The U.S. Census Bureau has considered three general sources for the acquisition of mailing lists from which initial address registers are printed: developing lists from field canvass operations, buying lists from vendors, and updating its lists from previous censuses.

The second major process in address register development after list acquisition is the updating process. The AR update operations are designed to improve the completeness (or coverage) of the residential address inventory. Basically, updating operations verify the ARs by (1) using enumerators to check addresses in the field or (2) using postal carriers to check Census Bureau lists against addresses on their delivery routes.

Finally, the delivery of census questionnaires is controlled by address registers. Census questionnaires can be sent to addresses using the postal service or delivered by an enumerator who takes the questionnaire to addresses in the AR and concurrently updates the address register. Additionally, censuses can be conducted without compiling pre-enumeration address registers. In 1970 and 1980, the United States conducted the census in non-mail areas by listing addresses concurrently with actual enumeration.

This paper will compare methods of address register compilation, methods of updating and refining AR completeness (coverage), methods of delivery and enumeration, and summarize AR usage.

Within each of these topics, a comparison will be made for different methods with respect to operational flow and AR processing, coverage gain, cost, and advantages/disadvantages.

## Notes:

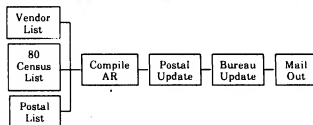
- A flow chart is presented in the text for each method of list acquisition, updating, and delivery to show how the particular method fits into census processing. The Appendix is a composite of these charts that illustrates the many address register development possibilities.
- Coverage and cost data in this paper come from widely varying sources and circumstances. Qualifications for these data, such as variance levels and sampling information, are beyond the scope of this report. Generalization to national census of the coverage and cost data given here is not recommended.

## Acquisition of Precensus Lists

The U.S. Census Bureau has considered four methods of address register acquisition:

- purchasing lists from private vendors;
- purchasing lists from the U.S. Postal Service;
- updating registers from the previous census;
- developing lists from field canvass operations.

In preparation for mailing, the operational flow of the first three methods can be simply represented as follows:



Address register compilation actually involves grouping of units on the lists into geographic areas (geocoding), resolving cases that do not geocode, and printing of address information in address registers. The ARs are then ready for the updating process.

Notice that between the list acquisition phase and the mail-out/mail-back operation, two updating procedures are undertaken: updating of addresses by the postal service and also by the Census Bureau. A basic goal of developing the address lists for a mail census is attaining consistency between the postal and Census Bureau's perceptions of addresses. That is, a mailing address that can be delivered by the postal service should be identifiable to an enumerator to locate the unit for interviewing.

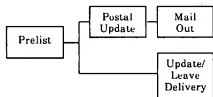
Coverage and costs of vendor, postal service, and (updating) 1980 census lists were recently compared in our Address List Compilation Test (ALCT).<sup>1</sup> For one urban area, the original list coverage and cost findings were as follows:

List	Coverage: Valid Units	Cost: Per Unit
Vendor	48,640	\$0.05
Postal Service	49,988(2.8%)	0.97
1980 Census	50,752(4.3%)	0.15

Two notes:

- The ALCT was carried out in 1984. The 1980 census list would be more outdated if used for 1990.
- As mentioned, the coverage figures given here are for the original lists. They do not reflect improvements which would be gained by updating techniques. (See the following section.)

The fourth list acquisition technique is a field canvass of areas by Census Bureau personnel. In this technique, the geographic area is delineated and an enumerator canvasses the area recording, by hand, address information in the address registers. (This information is later keyed for computer processing.) In 1980, this operation, called Prelist, was performed in areas for which no vendor list was available.



Questionnaire delivery in prelist areas may be accomplished in two ways -- by mailing or by a Census Bureau "update/leave" delivery. (This method will be addressed later.)

One measure of prelist and vendor list coverage is the address add rate found during the subsequent postal update operation. In 1980, there was no significant difference between postal update add rates for prelist areas and areas for which a vendor list had been purchased.<sup>2</sup>

A prelist operation is rather expensive. In 1980, the per unit cost of prelisting was \$1.39.<sup>3</sup>

It should be noted that in 1980 prelist was not subjected to an additional postal check, the advanced post office check, used for areas covered by vendor files. The advanced post office check will be explained below.

#### Advantages/Disadvantages Between Precensus List Acquisition Methods:

Costs seem to be the most pertinent factor in the U.S. list acquisition experience.

- In the ALCT the compilation of the postal list costs more than compilation by other methods.
- Use of address register listings from a previous census is a promising approach if continuous updating costs between censuses did not prove to be prohibitive.
- The need for field canvassing (prelisting) is apparent when no other list is available or adequate.

#### Updating of Precensus Lists

Precensus lists are updated in the operations described below. Updating serves two purposes:

- improving deliverability of addresses for mail censuses;
- checking the coverage of an initial list.

As mentioned earlier, a problem with precensus address registers is that the Census Bureau and postal service's perceptions of addresses sometimes differ. Furthermore, other problems associated with address lists such as vacant units

<sup>1</sup> Franklin, D., J. Dinwiddie, and M. Lueck, July 11, 1985. "Results and Analysis of the Urban Address List Compilation Test", Memorandum from Charles D. Jones to Susan M. Miskura.

<sup>2</sup> Whitford, D. and K. Thomas, June 30, 1983. "Post Office Effectiveness", 1980 Census Preliminary Evaluation Results Memorandum No. 52.

<sup>3</sup> No author, December 3, 1982. "1980 Decennial Census Costs".

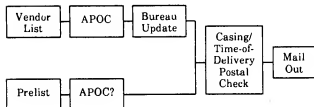
becoming occupied, new construction, demolitions, and ill-defined apartment numbers in multi-unit buildings necessitate the updating of address registers after they are initially compiled.

Three different scenarios have been used by the U.S. Census Bureau to update address registers. They are divided below into postal updating and field verification updating procedures.

### Postal Updating

The Advanced Post Office Check (APOC) was used in the 1980 census as a first check of lists which were purchased from private vendors.

This update occurred in the summer of 1979. Such a check can be fitted into a census operational flow as follows:



Note the question mark beside APOC for prelisted areas. APOC was not performed in prelisted areas in the 1980 census. But this type of early check of the deliverability of prelisted addresses is being tested for 1990.

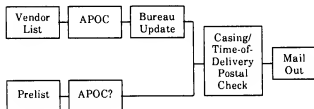
During the APOC, an address card was produced for each vendor address and sent to the post office. Residential addresses, to which the post office delivered mail, that were not included on the Census Bureau files were recorded on "blue cards" by the postal carriers. These were geocoded (if possible), checked for duplication and, if valid, added to the address registers.

The efficacy of this APOC operation is demonstrated by its coverage and costs. APOC added 5.5 percent of the total number of housing units eventually enumerated in the areas where vendor lists had been purchased. Ninety-three percent

of these were found to be occupied.<sup>4</sup> The APOC costs were not immediately available on a per unit basis but rather on a per unit added basis. Each address added by APOC cost the Census Bureau \$3.49.<sup>5</sup>

Just before the U.S. Census Day (April 1) when the census questionnaires were sent to the post office, two additional postal update operations were performed, the casing and time-of-delivery (C/TOD) checks. In 1980 the casing check was undertaken about three weeks prior to Census Day. Again, "blue cards" were returned to the Census Bureau for addresses without questionnaires. These were added to the address registers after a check of their validity. Then again at the time-of-delivery of the questionnaires, the postal carrier was asked to note any living quarters not receiving a census mailing package. These addresses were checked against the address register and added if valid.

The casing/time-of-delivery checks fit into census operations as shown by the following chart.



The casing and time-of-delivery (C/TOD) checks were carried out in all mail-out/mail-back areas of the country. Coverage improvement for the C/TOD checks was estimated to be 3.4 percent of all enumerated households in mail areas. Eighty-nine percent of these added units were found to be occupied.<sup>6</sup> On a per unit added basis, the cost of the casing and time-of-delivery was \$2.56.<sup>7</sup>

### Census Bureau Updating

Recalling the problem of attaining consistency between postal and Census Bureau

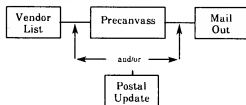
<sup>4</sup> Whitford, D. and K. Thomas, June 30, 1983, op. cit.

<sup>5</sup> Thompson, J., August 28, 1984, "Preliminary Summary Results from the 1980 Census Coverage Improvement Program Evaluations", 1980 Census Preliminary Evaluation Results Memorandum No. 85.

<sup>6</sup> Whitford, D. and K. Thomas, June 30, 1983, op. cit.

<sup>7</sup> Thompson, J., August 28, 1984, op. cit.

perceptions of addresses, the Census Bureau conducted its own updating procedure in January through March 1980. This was scheduled to occur between the two postal updates described above. This in-house update was called the Precanvass operation. In general, precavass fits into the operational flow as follows:



During the precavass operation, a Census Bureau employee used the (post APOC) address register to dependently canvass the housing inventory within the geographic area covered by the register. The employee recorded changes, additions and deletions of addresses, further identified any necessary physical location descriptions, and recorded the number of units in multi-units. For 1990, the Census Bureau is testing a unit-by-unit precavass in which each apartment designation in a multi-unit is listed in the address register and is individually checked.

After precavass, changes in address information are captured for the next copy of the address register and mailing pieces are produced which reflect this updating of the AR.

Coverage gains in the 1980 census from the precavass operation are hard to extract since many operations were happening simultaneously in census offices just before Census Day. But it is estimated that precavass was solely responsible for adding about 2.36 million housing unit listings to the ARs. The cost was about \$5.00 per added unit.<sup>8</sup>

#### Advantages/Disadvantages Between Updating Techniques:

Timing of updating operations and coordination of Census Bureau and postal versions of addresses seem to be the primary

concerns for address register development through updating methodologies.

- The APOC is used to update vendor lists soon after they are purchased, getting an initial check preparing them for final in-house and postal updating.
- The precavass operation helps resolve any disparity between mail and physical location addresses thus preparing the Census Bureau for post-enumeration followup activities.
- The casing/time-of-delivery checks occur when the post office actually has the questionnaires ready for delivery. For this reason it seems reasonable to have a post check at this time also.

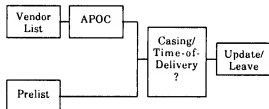
#### Delivery Methods

The various methods of delivering questionnaires to housing units include those involving precensus address lists (compiled in address registers) and methods where address registers are created concurrently with visits by enumerators. Mail-out/mail-back methods have been discussed above. In this section the use of an update/leave procedure and a list/leave procedure will also be explained. Coverage and cost analyses performed by the U.S. Census Bureau for these methodologies will be reviewed.

#### Update/Leave Delivery

Basically, an update/leave procedure involves Census Bureau employees **updating** address registers (compiled before the census) at the same time as questionnaires are being **left** at all housing units. An experimental update/leave scenario was undertaken during the 1980 census.

Operationally, it is described as follows:



<sup>8</sup> Fan, M. and J. Thompson, October 22, 1984. "Evaluation of the 1980 Census Precavass Coverage Improvement Operations", 1980 Census Preliminary Evaluation Results Memorandum No. 92.

Simply stated, this experimental procedure differed from operations throughout other mail areas in that:

- no prec canvass was performed and the address register was updated during delivery;
- reminder cards were produced and sent to addresses in update/leave areas.

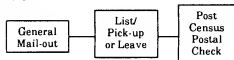
The experimental design for this 1980 test involved five pairs of census district offices paired by similar characteristics. In one office of each pair, update/leave delivery was substituted for prec canvassing (if not a prelist area) and mail delivery. Results indicated that even though census questionnaires were personally left at dwellings by enumerators in the update/leave offices (affording an opportunity to urge people to fill out their form completely), no significant differences in household coverage were discerned.<sup>9</sup> Update/leave methodology was more expensive than mailing forms to addresses.

#### List/Leave Delivery

In "conventional" census offices, where no precensus address list was purchased or compiled, 1970 and 1980 census practice was to:

- send questionnaires to households in a general (no address on the questionnaire) mailing;
- assign an enumerator to:
  - canvass designated geographic areas on or after Census Day;
  - list all housing units found in these areas;
  - pick up completed forms, leave a blank form, or interview the household at that time;
- conduct a postcensus address updating procedure where the post office is asked to check the addresses obtained at the time of enumeration.

The flow of these operations is represented below.



The coverage of this delivery method was directly compared to mail delivery methods in an experimental program during the 1970 census, the Mail Extension Test. In this experiment, five census district offices that had been designated to be conventional offices were redesignated mail-out/mail-back offices. They were paired in the experimental design with five other conventional offices.

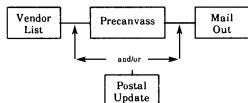
Results were as follows:

- There were no significant coverage differences between the conventional and mail offices.
- In the precensus list areas, the numbers of enumerator deletions and postal additions during updating were positively correlated with coverage error measures.
- For population data requested of everyone, mail areas had significantly more complete data.
- There was no significant difference in the cost of the two procedures.<sup>10</sup>

It should be emphasized that this 1970 test was undertaken in every rural area of the country. The Census Bureau used this list/leave approach in areas with many of these same characteristics again in 1980.

#### Postal Delivery

Address register acquisition and updating procedures for areas of the United States where a precensus address register is compiled have been described above. The basic flow for mail delivery areas is shown again as follows:



<sup>9</sup> Bailey, L. and P. Ferrari, June 15, 1984. "1980 Census Update List/Leave (ULL) Household Roster Check - Preliminary Report", 1980 Census Preliminary Evaluation Results Memorandum No. 70.

<sup>10</sup> U.S. Bureau of the Census, June 1973. **Results and Analysis of the Experimental Mail Extension Test, 1970 Evaluation and Research Program PHC(E).**

Comparison of the mail and list/leave approach is given above (the 1970 Mail Extension Test). Additional coverage information will be forthcoming in a report on the Post Enumeration Program for the 1980 census. It is designed to measure coverage for all parts of the United States and is about to be published. However, the confounding factor that geographic/demographic characteristics of list/leave and mail census areas are quite different would limit the usefulness of this coverage comparison. Similar limitations accompany any cost comparison between the two types of areas. But cost figures for the 1980 census are available. For all but one of the Census Bureau's regional offices, field costs per unit in conventional (list/leave) areas of the country exceeded those in mail areas.<sup>11</sup>

#### **Advantages/Disadvantages Between Delivery Methods:**

The advantages and disadvantages of using a list/leave or mail-out/mail-back approach to census-taking are not clear. In one comparison in rural areas, no noteworthy differences were discerned.

Historically, in the United States Census Bureau's experience, it seems that rural areas and with sparse population lend themselves to a list/leave approach, whereas the more urban populations are better enumerated using address registers compiled before a census is undertaken.

#### **Use of Address Registers**

The preceding discussion follows the early life cycle of address registers through various compilation methods, improvement

strategies, and roles in various questionnaire delivery techniques.

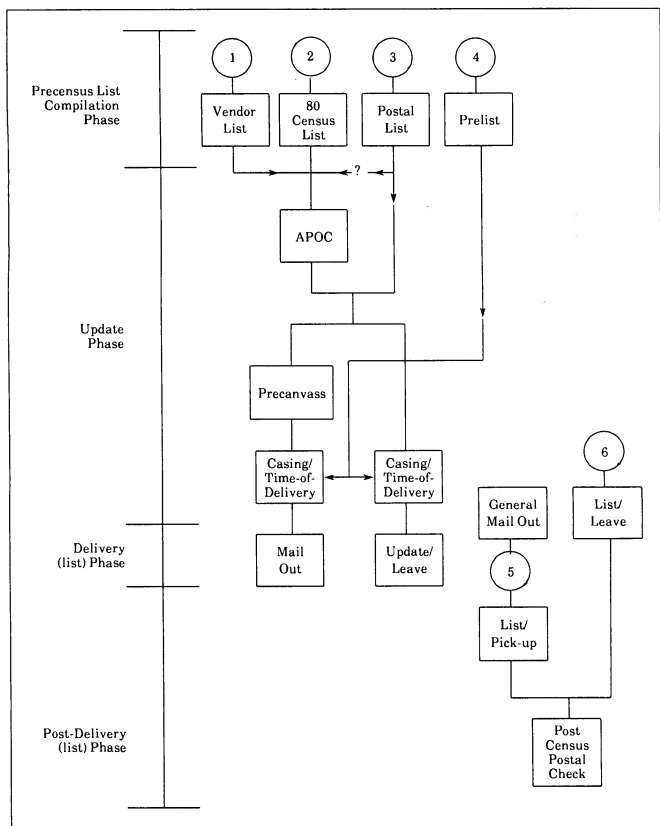
To reiterate, address registers are used to record information for living quarters and to control certain aspects of the field enumeration process. The information recorded may include any or all of the following: geographic codes, address (street name and house number, apartment designation, city, state, zip code), householder name, and physical location description. These data and other information entered in the address register can be used to control or provide:

- workload estimates used in planning staff size, logistics and supplies, budget preparation and so forth;
- correction, additions, and deletions of addresses;
- data for evaluations (e.g., address coverage gain from an updating procedure);
- receipt/non receipt status of questionnaires to allow planning for followup activities;
- preliminary data on population and housing counts;
- quality control of content sample collection.

In summary, address registers provide a means by which the field enumeration phase of a census can be controlled. The management and control options inherent from address register usage are many.

<sup>11</sup> Computer print-out, August 31, 1983. "Table III. 1980 Census Field Operations Costs/Regional Office Costs by Type of District Office".

# APPENDIX: SIX ADDRESS REGISTER DEVELOPMENT POSSIBILITIES





# APPLICATIONS OF AN ADDRESS REGISTER IN THE CANADIAN CENSUS

DON ROYCE

CENSUS AND HOUSEHOLD SURVEY METHODS DIVISION  
STATISTICS CANADA

## Introduction

One of the key activities in the conduct of the Census of Population is the creation of a list of all dwelling units in Canada. For 1986, this task will be done on an enumeration area (EA) basis by the Census Representative (CR) and the list of dwellings for each EA will be recorded in the Visitation Record (VR).

This list has a multiplicity of uses in the conduct of a census:

- (a) it helps to ensure complete and accurate coverage of the EA by recording a unique description of each dwelling;
- (b) it serves as a sampling frame for the CR to leave a longer questionnaire at every fifth household;
- (c) it serves as a record of the status of each dwelling during collection, e.g., date questionnaires dropped off, date questionnaires received back, follow-up required, name of person to contact;
- (d) it serves as a source of control information (e.g., person and dwelling counts) for subsequent processing of questionnaires; and
- (e) it serves as the source of interim dwelling and population counts.

As mentioned, this list of dwellings has traditionally been created by the Census Representative, using the Visitation Record. The listing operation is combined with the operation of dropping off the questionnaires. This combination has proven to be a very effective methodology, as it allows for person-to-person contact with respondents, for the determination of the number of questionnaires to leave in large households and in what language, for the CR to observe and record data such as dwelling type, and for a relatively efficient sampling methodology. This methodology ensures that the list is up to date, as it is created during a short period before Census Day. It is also believed to be a strong motivator for the CR in that the CR has responsibility for the complete enumeration of his or her EA.

Despite the success of this method in the past, however, it is my belief that some investigation of an alternative approach is warranted. For one thing, many of the advantages of the drop-off technique are declining in importance. Contact rates at drop-off are 50% and declining, we now have a bilingual questionnaire for 80% of households, and households themselves are getting smaller. Increased concern for personal security also means that physical access to some types of dwelling units is becoming more difficult.

We are also experiencing major changes in the ways in which our work-force is recruited. We may not be able to depend on the highly motivated CR for good coverage in 1991.

The fact that the list of dwellings is only in manual form not only means that it must be created from the beginning again each census, but, as described in the next few paragraphs, it also limits many other aspects of census-taking both in collection and processing.

The alternative discussed further on in this paper is that of a computerized household address register (AR). You will also find a description of an address register along with some possible applications, a description of previous research on this topic as well as a discussion on possible sources for an address register. The concluding part of this paper suggests a plan for future research and other considerations important in the development of an address register approach.

## The AR Concept and Potential Applications

The concept of an AR referred to in this paper is a file, in machine-readable form, of residential addresses for some or all portions of Canada. Each address would correspond to a dwelling unit, whether occupied or not, conforming to the census definition. Each address in the register would include the Census Geography coded to the maximum extent possible but at a minimum to the level of the enumeration area. For each address, certain supplementary information would also be recorded, for example, the type of dwelling unit and possibly the telephone number. However, the AR would not contain the names or any other personal information about the occupants of the

dwelling. It is anticipated that the AR would be maintained on an ongoing basis throughout the five-year census cycle. Additions, deletions, and changes to the file would be made on a continuing basis as the update information became available.

The major uses of such an address register might be as follows:

#### **Mail-out of the Questionnaire**

With a reliable AR, the census questionnaire could be mailed out rather than dropped off. This method is used by the United States for much of its population. Comparative cost studies done as part of the earlier research forecast savings of between \$40,000 and \$652,000 for the 1981 Census, and between \$569,000 and \$2.0 million for the 1986 Census (in 1976 dollars).

Most of the cost savings would come from the elimination of the drop-off phase of the CR's work, including drop-off training. In mail-back areas, since staff would only be required for follow-up, a CR could handle a large assignment, thereby leading to an overall reduction in the number of field staff required. This might also permit the hiring of better qualified people. As well, a number of checks such as the postal check and geographic field checks prior to census could be eliminated.

#### **Coverage Improvement**

Even if a mail-out census were not adopted, an AR could be used to improve coverage, reduce listing costs or both. One option would be to use the AR to replace the postal check, which has now become extremely expensive (\$1.4 million in 1981). Alternatively, the CR could be given a Visitation Record pre-printed with the addresses in the EA. He or she would then simply have to update the VR rather than create it. This has the potential for both lower costs and increased accuracy.

#### **Delineation of Field Assignments**

CR assignments (the same as or similar to existing EAs) could also be delineated on the basis of counts of the number of households from an AR. Availability of an AR could reduce or eliminate the need for geographic coverage checks prior to the census and could reduce the need for EA splits at census time.

#### **Use of Alternative Sampling Methods**

With an AR, the 2B sampling could be done in advance. By having more control over the selection of the sample, the potential for drop-off bias would be reduced. As well, more complex methods could be used, for example, varying sampling fractions, deeper stratification, or interlocking sampling. Increased efficiency in sampling might allow a reduction in the overall sampling fraction and, therefore, a reduction in respondent burden.

#### **Special Handling of Collectives**

By identifying collective dwellings on the AR, special field procedures could be put in place to ensure that these dwellings are handled properly. In fact, an AR already exists for such dwellings in the form of the Collective Dwelling Listing File which is created from the previous census. An AR would simply extend this existing file by supplying updating information for the intercensal years.

#### **Telephone Follow-up of Complete Non-response**

In the 1986 Census, the mail return before follow-up is expected to be between 80 and 90%. With 6 million households in mail-back areas, this still leaves between 600,000 and 1.2 million households requiring follow-up – the equivalent of between 10 and 20 Labour Force Surveys.

Considerable savings can be achieved by following up these households by telephone rather than field follow-up. In 1986, the CR will attempt to follow up non-response households by telephone whenever possible. However, the CR will be able to obtain a telephone number only when the name of Person 1 is obtained at drop-off. This is expected to be about 50% of the time.

An AR which also contained the telephone number for each address could increase this percentage considerably. Using the AR, we could prepare a "mini telephone directory" for each EA. A print-out, sorted by address and showing the telephone number, could be used by the CR to locate the telephone number for any address where a questionnaire was dropped off but not received back.

For this application, the completeness and accuracy of the AR are not extremely critical. If an address is missing, the CR would simply revert back to the existing procedure.

A test of this application is planned for the 1986 Census, in Ontario and Quebec, making use of a file already purchased from the telephone company. Using the postal code on this file, we plan to re-structure the file by 1986 EA and provide a print-out as described above to the CR.

### **Centralized Edit and Telephone Follow-up**

The pros and cons of centralized edit and telephone follow-up have been documented in earlier papers and are not repeated here. If, however, it were decided that such a methodology was desirable, an AR could serve as the basis for a file to monitor and control such an operation. The control file could be used for:

- (a) registering questionnaires as they are received from respondents;
- (b) recording the status of each questionnaire through the edit and follow-up operation;
- (c) automatically creating and controlling assignments for both telephone and field follow-up;
- (d) administrative functions such as production of MIS reports, calculating and verifying pay accounts; and,
- (e) a source of data for subsequent analyses of the operation.

In effect, the control file would be a computerized version of the current Visitation Record.

### **Control File for Census Processing Activities**

A control file based on an AR could be used not only in collection activities, but also in many of the subsequent processing steps. For example, the file would be used to control the coding and the capture of questionnaires. We could then immediately link the data from the "VR" (now computerized) to the questionnaire. Many of the steps in the current Head Office Processing operation could be streamlined with a machine-readable VR.

Any other census operations which involve access to the VRs would also be simplified.

In the long run, an automated AR could facilitate the integration of collection and processing operations. One could, for example, envisage a scenario where the questionnaires are mailed back and then immediately captured (perhaps after some minimal pre-grooming), edited by computer, and batched for follow-up. Since the questionnaires are already captured, follow-up could even be via Computer-assisted Telephone Interviewing (CATI). Once follow-up is completed, the questionnaires can be sent for edit and imputation. At the same time, interim population counts could be produced.

### **Enhancement to the Area Master File (AMF)**

An AR could potentially serve as a source of updates to the Area Master File. It could essentially extend the level of geographic coding from the level of the block-face to the level of the household. This would allow a method of direct geocoding of any household in any survey. This would in turn promote the implementation of the new Standard Geographical Classification (SGC) policy and would enhance the small area data capability of Statistics Canada. An extension of the level of coding of the AMF might also be of use to many of the external users of the AMF, such as municipal police forces.

### **Frame for Household Surveys**

In addition to these census applications, an up-to-date AR could serve as the sampling frame for a wide variety of household surveys, including the Labour Force Survey, post-censal or intercensal surveys linked to the census questionnaire, and independent ad hoc household surveys. An increased ability to conduct post-censal or intercensal surveys linked to the census data base could permit a reduction in the number of questions on the census itself. With an increased demand for data on specialized subgroups of the population, we might move in the direction of using the census as a screening vehicle with more in-depth questions contained in a follow-up survey. The usefulness of the AR for this purpose would be greatly enhanced if the telephone numbers were also available.

Even if the addresses themselves were not used for sampling, an AR could serve as an

alternate source of up-to-date counts of households for the design of ad hoc surveys or the update of continuing surveys such as LFS.

### Previous Research on Address Registers

In the mid-1970s, a series of research studies was carried out on the financial and operational feasibility of creating and updating an address register for major urban areas in Canada. A summary of these studies is given in Booth (1976).

The major application for the address register created in these studies was for a mail-out Census of Population. Consequently, the primary focus of the studies was on the coverage of the AR. Three measures of an AR's coverage relative to a perfect AR were defined as follows:

The % complete gives a measure of the under-coverage of the AR, the % accurate gives a measure of the overcoverage of the AR, and the % reliable gives a combined measure of quality.

In the first study, an address register was created for five cities and was updated for 18 months with data from a number of sources, including Canada Mortgage and Housing Corporation, building permits, and the post office. Table 1 shows the % complete rates in this study.

The major problem with the AR was in the location of subaddresses in the core of large urban areas. Checks showed that the majority of the undercoverage was due to older houses that had been modified to two- and three-unit dwellings. For the two smaller cities, however, the AR was superior for both singles and subaddresses.

% complete	= 100	- number of valid units not on AR	
		number of valid units on a perfect AR	* 100
% accurate	= 100	- number of invalid units on AR	
		total number of units on AR	* 100
% reliable	= 100	- (number of invalid units on AR + number of valid units not on AR)	
		number of valid units on a perfect AR	* 100

**TABLE 1. % Complete for Address Register (AR) Compared to Visitation Record (VR) for Five Test Cities**

City	Type of dwelling					
	Single		Subaddress		Total	
	AR	VR	AR	VR	AR	VR
Montréal	99.48	99.19	98.08	98.42	99.19	99.03
Toronto	99.77	99.45	89.96	98.42	97.15	99.40
Vancouver	99.63	99.35	97.90	99.72	99.10	99.46
Sherbrooke	99.61	99.22	99.47	98.08	99.55	98.77
St. Catharines	99.65	99.62	98.50	97.57	99.34	99.06

**TABLE 2. % Complete, Accurate and Reliable for Address Register (AR) Compared to Visitation Record (VR) for Two Groups of Cities**

	Montréal, Toronto and Vancouver		Sherbrooke and St. Catharines	
	AR	VR	AR	VR
% complete	98.64	99.31	99.45	98.91
% accurate	98.78	96.78	98.17	98.72
% reliable	97.43	95.97	97.59	97.62

Table 2 shows the % complete, % accurate and % reliable rates for two groupings of these cities.

Based on the first study, the AR does seem to be very comparable in quality to field listing. Coverage of the AR is worse in the larger cities but better in the smaller areas. In terms of accuracy and reliability, the AR is superior to the VR in the larger urban areas, while in the smaller urban areas the VR is just slightly better.

Subsequent studies concentrated on the question of whether an address register could be maintained from just one source, namely "Point of Delivery" sheets for Canada Post. These are sheets which contain every basic address where mail is delivered on each postal route. They are maintained in the local post office and are updated by the postal supervisor.

One of the studies was done for Trois-Rivières, one of the 1974 Census Test cities. The results for an address register maintained for 11 months using only Canada Post data are given below.

In this case, the AR is better than the VR in terms of all these measures - completeness, accuracy and reliability. The only area where the VR is better is in collective dwellings. Similar studies for a few other cities were carried out, but their results are not reproduced here. The figures are comparable.

In summary, the previous studies indicated that an address register could be created and updated with information from one source (Canada Post), and that the quality is in most cases at least as good as field listing. The one area of weakness seems to be in the identification of subaddresses in the larger cities, and measures to handle this would have to be devised.

Despite the apparent potential of an AR approach, however, the capital cost involved was considered to be problematic and the research did not continue.

#### Possible Sources for an Address Register

Since the earlier studies were carried out, there have been a number of developments which may make the creation and maintenance of some form of address register by automated means a viable proposition within the next five years. These include:

- the availability, or potential availability, of an increasing number of sources of up-to-date administrative data, such as records from telephone and hydro companies, Family Allowance, Old Age Security, and Revenue Canada files;

**TABLE 3. % Complete, Accurate and Reliable for Address Register (AR) Compared to Visitation Record (VR) for Trois-Rivières**

	Type of Dwelling							
	Single		Subaddress		Collective		Total	
	AR	VR	AR	VR	AR	VR	AR	VR
% complete	99.16	98.78	98.54	92.78	88.51	93.10	98.93	97.07
% accurate	99.10	99.42	96.88	90.06	100.00	100.00	98.47	96.79
% reliable	98.26	98.21	95.36	82.54	88.51	93.10	97.40	93.79

- (b) the universal use of the postal code on these files, which, combined with Geography Division's Area Master File/Postal Code link, facilitates the restructuring of these files into Census Geography by automated means;
- (c) improved record linkage methodologies which could be used to unduplicate multiple files; and,
- (d) the declining costs of storing and working with large data files.

Earlier this year, a number of potential sources of addresses were briefly investigated. The results are summarized below:

#### **Canada Post "Point of Delivery" Sheets**

**Contents:** Address, postal code, business/residential indicator, number of mail receptacles for each address, other information related to the amount of work involved in delivering mail.

**Coverage:** Based on studies ten years ago, comparable to or better than field listing. Includes only addresses receiving home delivery, but other records are apparently available for these exceptions.

**Availability:** Only one copy exists and is kept at the local Postal Installations. There may or may not be photocopiers at these offices.

**Frequency:** To be determined.

**Comments:** Canada Post would also like to store these lists in machine-readable form.

Maps showing the letter carrier routes are also kept in the local office.

#### **Telephone Company Files**

**Contents:** Subscriber's name, address, postal code, business/residential indicator, and telephone number.

**Coverage:** Subscribers with published telephone numbers.

**Availability:** Bell Canada (most of Ontario and Quebec), to be negotiated elsewhere.

**Frequency:** Quarterly.

**Comments:** Address information is from billing system and should therefore be of high quality.

Address of equipment installation is also available.

#### **Municipal Assessment Lists**

**Contents:** Address, postal code, assessment, category (regular, cottage, farm), type of dwelling.

**Coverage:** To be determined (in theory should be good).

**Availability:** Available at cost in Ontario, to be determined for other provinces.

**Frequency:** Annual.

**Comments:** Assessment by municipality may be a by-product.

#### **Hydro Companies**

**Contents:** Address, postal code.

**Coverage:** Problems with block metering of apartment buildings.

**Availability:** Central hydro utilities exist in all provinces except Ontario, to be determined.

**Frequency:** Potential for quarterly.

**Comments:** Address information is from billing system and should therefore be of high quality.

#### **Family Allowance**

**Contents:** Address, postal code.

**Coverage:** Persons with children.

**Availability:** Statistics Canada to receive file from Health and Welfare shortly.

**Frequency:** Believed to be monthly.

**Comments:** Address information used for cheque delivery and should therefore be of high quality.

#### **Revenue Canada Taxation**

**Contents:** Address, postal code.

**Coverage:** Individual tax filers.

**Availability:** Statistics Canada already has.

**Frequency:** Annual.

**Comments:** Problems exist with timeliness, multiple filers per address, and addresses being tax discounters.

#### **Area Master File**

The Area Master File itself could be used to generate an address register. For each block-face, the AMF contains the street name and the address range, e.g., 2 to 54. One could generate addresses by assuming there was a dwelling at every second number within the address range, e.g., 2, 4, 6, ..., 52, 54.

**Contents:** Address, postal code, Census Geography.

**Coverage:** Same as AMF (most cities over 50,000 and increasing).

**Availability:** Statistics Canada file.

**Frequency:** Continuous.

**Comments:** Overcoverage of addresses would likely be a major problem.

#### **Labour Force Survey Apartment Frame**

**Contents:** Address of building, postal code, 1981 Census Geography, number of floors, number of units.

**Coverage:** Most CMAs, all buildings with 30 or more units and 5 or more floors.

**Availability:** Statistics Canada file.

**Frequency:** Continuous.

**Comments:** Updates to the frame can be up to 12 months late.

#### **Visitation Records from 1981 Census**

**Contents:** Address, agricultural holding indicator.

**Coverage:** All dwellings in 1981 Census.

**Availability:** In paper form only, would have to be data captured.

**Frequency:** Every five years.

**Comments:** Could serve as a source for creating an AR, but other methods would obviously have to be found for maintenance. Previous studies have indicated data capture to be very expensive.

In addition to the sources described above, there are several other sources of addresses which are worth investigating, including provincial driver's licence and medicare files, commercial sources such as those used by the U.S. Bureau of the Census, and voters' lists compiled by federal and provincial electoral officers.

#### **Further Research and Outstanding Issues**

A number of steps are proposed for further research into this area:

1. Carry out a more comprehensive study of the various sources and approaches to the construction and maintenance of an address register.

The various sources mentioned previously would be investigated in further detail, as would record linkage methods appropriate to the unduplication of addresses. This would also include an investigation of a possible joint venture with Canada Post to automate their lists.

2. Become more familiar with the experiences of other countries who use an address register and/or mail-out approach for their censuses.
3. Update the cost comparison exercises done in 1976 to better estimate the potential cost

savings with an address register - mail-out census methodology. A cost comparison exercise for the Labour Force Survey should also be done.

4. Construct a pilot register using an administrative data/record linkage approach.
5. Evaluate the quality of the pilot register in a post-censal study and with the ongoing LFS. Evaluation would focus on both the coverage of the AR and on the quality of data items on the AR. This study could be combined with other field testing for the 1991 Census.
6. If the mail-out methodology appears to be feasible and promises sufficient cost savings, conduct a field test to measure respondent and interviewer reaction, to estimate response rates, and to detect any problems with the methodology.
7. Update cost comparisons based on field test of mail-out methodology.
8. Test the use of the AR as a frame for the Labour Force Survey.

A proposed schedule for testing related to census is shown in Table 4.

Because a number of programs within Statistics Canada stand to benefit from the availability of an address register, it is proposed that a number of areas participate in this work. These include the Small Area Data Program, Administrative Data Development Division, Census and Household Survey Methods Division (both subdivisions), Survey Operations Division, Geography Division and Informatics Services and Development Division.

There are also a number of concerns which would have to be addressed before the applications mentioned previously could be implemented.

First, the use of a mail-out methodology would make us more vulnerable to errors, interruptions or cost increases in the postal delivery service. With the present method, for example, a post office strike would severely disrupt collection, but not totally destroy it, since the CR could resort to pick-up of questionnaires if necessary.

Second, what would be the public perception of an address register? The public may not distinguish between this and a population register with all of its "big brother" connotations.

Third, a fully successful AR would require more geographical standardization than now exists. At the present time, the Census Identification (Electoral District, Enumeration Area, and Household Number) for the same dwelling unit changes from one census to the next. Some form of block program would be required which would maintain the stability of the geographical identification down to the block-face level.

Fourth, the scope of the AR would have to be determined. At present, it would appear to be much more feasible in urban areas than in rural areas. However, it is in such areas where the potential cost savings are likely higher due to elimination of travel costs for drop-off.

Fifth, there is the question of how dependable the various sources of information needed to maintain the AR would be. For example, Canada Post has recently begun reducing the extent of home delivery of mail as a cost-cutting method. As a result these addresses may not be available or may be difficult to obtain.

Finally, there is the question of how we might reduce the risk in making a wholesale change-over from our traditional approach to an approach based on an address register. We would, in fact, be converting from an "area frame" approach to a

**TABLE 4. Milestones for Address Register Research**

Activity	Date
1. Study of AR construction and maintenance	September 1986
2. Familiarization with experience of other countries	October 1986
3. Update cost comparisons from 1976	June 1987
4. Construction of pilot register	March 1987
5. Post-censal test of coverage of pilot AR	June 1987
6. Test of mail-out methodology	June 1988
7. Update cost comparisons based on mail-out test	March 1989

"list frame" approach. Previous research, while encouraging, has been based on a relatively small number of cities.

One possible solution is to create and maintain a full scale national AR for 1991 without actually using it in data collection. The quality (completeness, accuracy and reliability) of this AR would be evaluated immediately following the 1991 Census through a 100% comparison and field check. Because this approach would compare two independently created lists, the evaluation of the AR's quality is likely to be quite reliable. However, we would get no information on problems of using the AR in the field. Such a comparison would also be difficult and expensive to implement.

A second approach would also maintain the current collection methodology (using drop-off), but the CR would be provided with the AR for

his/her EA as a starting point for listing dwellings. Updates made by the CR would be data captured and the quality of the AR evaluated by comparing the original to the updated version. In effect, this would build the field check of the AR into the 1991 Census itself. However, since the verification of the AR is a dependent verification, there is some risk that the quality of the AR could be overstated. On the other hand, we would get valuable experience with the use of an AR in the field and there may be the potential for some cost savings.

Whatever the approach, it is clear that the full potential of an address register would only be realized with a change-over to mail-out of the questionnaires. Such a change would have to be very carefully tested and proven, but the fact that the U.S. does use this method successfully is encouraging. Whether we wish to move in this direction for 1991 is now up to us.

---

#### REFERENCES

Booth, J. K., 1976. "A Summary Report of All Address Register Studies Completed to Date", Statistics Canada Report, E-414E.

Booth, J. K., 1976. "NAR cost comparisons - Present Worth Approach", internal Statistics Canada memorandum.



## **Résumé of the Question Period**

*During the question and discussion period several questions and issues were raised.*

### **Coverage**

*The first issue was whether international comparisons of the results of coverage measurement studies could yield insights into the reasons for under-coverage, for example, what types of areas or types of population have particularly high rates of undercoverage. Such comparisons have been attempted, but are severely confounded by differences in the techniques countries use to measure coverage error. For example, the U.S. gets widely different estimates from its demographic analysis and post-enumeration surveys. The only way to try to remove these "method" differences would be to run a controlled experiment.*

*It was pointed out that different techniques may also be more suitable for certain populations or certain levels of aggregation.*

*A third point raised was that the U.S. and Canada put more emphasis on coverage measurement and the U.K. concentrates more on measuring response error, although all these countries measure both.*

*A final and more general point was whether the trade-off between quality and timeliness was different for censuses, which are used over a 5- to 10-year period, than for monthly, quarterly and annual surveys. It was pointed out that in practice it was not possible to decide to take more time in order to improve quality part-way through the census process, and that census results need to be timely since the population can in fact change rapidly.*

### **Edit and Imputation**

*The major point raised in the discussion on this topic was that statistical agencies are subject to questions about the methods used to edit and impute data. Furthermore, the trend is towards increased user sophistication and awareness of such methods. Examples were given of such occurrences with language data in Canada, and with race and Hispanic origin questions in the U.S. A number of different possibilities for the future are foreseen. Statistical agencies may have to negotiate how data are edited and imputed in a manner similar to which question wordings are negotiated. We may have to take a stance of not attempting to impute data with the implications this has for multiple sets of figures derived from the same questions. At the same time, we must*

*take account of the fact that comparability over time is an important consideration for many users.*

*A second item of discussion was whether a strategy of releasing unedited data first followed by edited data at a later date would be advisable. There was some feeling that such a strategy would delay the release of the final data. Whether having better timeliness for unedited data at the expense of worse timeliness for final data was a valid trade-off is an open question.*

### **Address Registers**

*Several questions related to the scope of the proposed Address Register were raised. The possibility of extending it to cover farms, perhaps using the Farm Register as a base, was mentioned. It was felt that for 1991 it was most likely that it would be restricted to urban areas.*

*The U.S. commented on some of the considerations they had made in deciding to use an address register and the mail-out/mail-back method. It was emphasized that address registers are expensive, but that the U.S. had decided to go in this direction for the gains in coverage and control. As well, the size of the U.S. population rules out being able to send an enumerator to every dwelling, making a mail census a necessity. Third, the U.S. has the advantage that in many areas vendor tests are available which are relatively inexpensive and of good quality; the same may not apply in Canada. Fourth, although the initial cost may be high, it should be noted that the cost can be amortized over several censuses. With these caveats, however, further research into address registers by Statistics Canada was seen as worthwhile.*

*The U.S. also responded to a question on whether they will integrate their Address Register into the TIGER system for 1990. The answer was no, largely because the TIGER system was a major undertaking in its own right and adding this additional step was seen as too risky at this time.*

*The issue of privacy was also raised. For an address register to be useful, the telephone number would be important. However, this was close to becoming a population register. Research would be needed into public acceptability of an address register. It was pointed out that in the U.S. such address registers are treated as confidential information.*

*Finally, a third approval to testing an address register in 1991 was suggested. A test area could be chosen where an address register and mail-out/mail-back would be attempted, with a contingency plan in place in case the test failed.*



## **SESSION: THE ROLE OF RESEARCH AND TESTING**

Chairperson: E. T. Pryor  
Director General  
Census and Demographic Statistics

Friday, October 11, 1985



# PLANS FOR THE 1991 CENSUS IN THE UNITED KINGDOM

DAVID PEARCE

OFFICE OF POPULATION CENSUSES  
AND SURVEYS  
UNITED KINGDOM

## Introduction

*Before starting, I would like to say how much I have enjoyed the last three days. It is good to meet people of other countries that are involved in taking censuses. It has been very stimulating. Briefly, I will just list some of the benefits that I derived from this conference.*

*First of all, there is the issue of the time-scale involved in planning the 1991 Census of the United Kingdom. Second, it provided me with some comfort because the problems we face are also faced by other census-taking countries. Third, it reinforced the fact that even if we are challenged by new technology and the excitement that changes bring on, we are still mindful that the census is concerned with economic, timeless and accuracy factors. Finally, it is important that the questions asked will not only be considered from a technical point of view but also be examined in such a way as to be acceptable by the public.*

*In terms of giving an overview of the censuses in the U.K., I would like to look back at lessons we have learned between 1971 and 1991. Following this, I would like to talk about the current position of the U.K. and where we are going from there. There are general issues such as maintaining the public confidence in accepting the census, or maintaining the security of the system and the public perception of it, or balancing the needs of users in the area of technology. These are very broad issues that I will not address in order to concentrate on more specific issues.*

## Census Experiences from 1971 to 1991

*In 1971, one of the problems was that we had an inflexible tabulation program that was inefficient and expensive in computer time. There was a huge tabulation program that tended to be uncoordinated and which became almost uncontrollable.*

*In 1981, we tried to improve the approach to the tabulation program by using data streams which created data series. After, we established key dates*

*in order to release figures to support a rate support grant, which is a financial distribution of money to local authorities. Then, we developed a timetable and identified the critical points for planning the output.*

## 1981 Output Program

*How successful were we in 1981? First, the support grant figures were provided by the required date. This represents a major achievement. Also, a series of county monitors which gave preliminary results for each county were produced between the period of October 1981 to June 1982. Improvements were realized in the published program compared to the 1971 Census. Publications were released from October 1981 to June 1982 and the volume on country of birth was published in February 1983. For the 1971 Census, the publications were released from May 1972 to November 1973, and the volume on country of birth was published in December 1974. The 1981 situation was a considerable reduction in the time-release scale.*

*Even with the success of the 1981 output program, some improvements could be made in other areas for the next census.*

## Collection Procedures

*One of the areas that should be looked into is the underenumeration. Although the overall rate was only 0.5%, it varied considerably. For example, in inner lands it was 2.5%. The major sources were: first, persons missed in enumerated households and second, households that were classified as absent on census night when in fact one or more persons were living there. The estimate of wholly absent households, as determined by the post-enumeration survey, showed a much lower rate than the enumerator had assessed during the enumeration. By looking at the collection procedure, it should be possible to review the way absent households are identified and the procedural checks used to check the enumerator's work.*

## Quality of Answers to Questions

*The U.K. uses sophisticated procedures to verify the quality of answers to questions. The error rate for some questions revealed a misunderstanding of certain questions. For example, the gross rate of errors for the question on the number of rooms was nearly 30%; another was the employment state which was around 10%. In general, it appears that the public and the users are prepared to accept inadequacies in the figures. However, it is still necessary to check the error rate of specific questions in order to assess the relevance of the information collected.*

## Population Bases

*In 1981, preliminary counts were produced using the enumerator record book which is a manual count (the preliminary count was 49 millions while the final count was 49.15 millions).*

*Then, two usual resident bases were created. The first base was the present/absent usual resident base. This base did not include wholly absent households on census night and thus yielded a lower population count (48.52 millions) than the transferred usual resident base of the country (49.15 millions) which did include total absent households.*

## Current Position on Planning the 1991 Census

*Some guidelines for the planning of the 1991 Census have been developed. The first step was to set up the policy committee which is composed of deputy directors, including interested divisions. This committee has already established some broad criteria for the 1991 Census. First, it was established that the 1991 Census should be a simple full census. Second, we should start upon the consultation process immediately. Third, the timetable for the published data output should be reduced. Fourth, the level of confidentiality and security should be maintained as it was for the 1981 Census.*

*In addition, we have agreed on several other aspects such as a broad timetable on consultation. We have also agreed that we should have a testing program which would include a major test in 1987 and a test of the ethnic question. Again, in 1987, we should publish a white paper on the 1991 planning census, to be presented to Parliament. This paper should also be available to the public.*

## Development of Project Groups

*The development of project groups permits the identification of specific issues that should be*

*examined in order to be considered for the research and testing program. At this point in time, six groups have been set up. The first group is reconsidering the geographic issue for the enumeration procedures. In England, in the late 1970s, the mapped postal unit boundaries represented about 1.6 million postal code units. Each unit is composed of 15 to 17 addresses. The digitizing of postal units would not only be beneficial in planning the enumeration districts but also in the output program. In addition, a lot of statistics are related to the postal code units. The ability to derive statistics from different sources of compatible areas would be facilitated involvements. Finally, the market for postal code information certainly exists but the major difficulty is the cost of such research. Again, it is difficult for census planners to obtain funds for a research and testing program, especially when the plans for this census are still at an initial developmental stage.*

*The second group is investigating the processing strategy that should be adopted. First, the problems involve managing a very large district. Another critical aspect is the choice of a hardware environment because of the time involved in testing and acquiring such a system. A new hardware environment raises several questions such as: What kind of system? Should we think about a dual system? Should we have a distributed or a centralized edit system? Should we decide to have an OMR, OCR system for basic counts and controlled keying for the other counts? What parts of the clerical process should be automated? These are essential questions that will need to be addressed in the near future.*

*The third group is looking at the feasibility of having an ethnic origin and language test. In the 1981 Census, we did not include a question on the ethnic language variable. However, we are committed to the testing of ethnic and language questions for 1991. There are three issues on the ethnic language variable. The first is to show that the question will work, the second is to find a question which is acceptable, and the third is to ensure that there are justifications for including an ethnic and language variable in a census.*

*The fourth group has been set up to identify population bases. The main issue for population bases is related to usual residents, and two possibilities can be raised. First, could we transfer visitors? Second, and perhaps more feasible, could we obtain data from wholly absent households? However, we must consider the legal aspects concerning the return of two forms.*

*The fifth group is studying the enumeration procedures, and I will not comment further on this.*

*Finally, the sixth group is looking at economic activities. First, we need to update the occupation classification. Second, there is a need to verify if the updating of the classification will reflect the changes in employment that have taken place in the last ten years. Third, the representation of women in the occupational classification has been*

*criticized and needs examination. Fourth, we need to consider international occupational classifications and other classifications in the country. At last, it is necessary to look at the way in which we aggregate the classification.*

*Thank you very much.*



# MAJOR ISSUES IN THE 1990 U.S. CENSUS OF POPULATION AND HOUSING

PETER A. BOUNPANE

DEMOGRAPHIC CENSUSES  
U.S. BUREAU OF THE CENSUS

## Introduction

The next U.S. Census of Population and Housing will be conducted as of April 1, 1990, and will mark the 200th anniversary of census-taking in the United States. Although 1990 is 4 1/2 years away, planning has been underway for some time at the Census Bureau. With the number of decisions to be made and the long lead times required, early planning is necessary.

When one considers that we are about to observe the bicentennial of census-taking in the United States, the next census takes on added importance: We want to take a census that will be worthy of our long heritage. The 1990 census will be the 21st in an unbroken chain since 1790 and will produce data to carry us up to the 21st century.

Progress comes from building on the experience of what worked well and what worked poorly. As a start, therefore, we made a thorough examination of the 1980 census. On balance, the 1980 census was a success and had several major accomplishments:

- Estimates show improvement in coverage over the 1970 census.
- We delivered the counts for reapportionment and redistricting by the legally mandated deadlines.
- The public information and outreach programs were highly successful, helping us achieve an 83-percent mail return rate and good coverage.
- We produced more data, particularly for small areas.

This is not to say there were no problems with the 1980 census. Some of the major problems the Census Bureau faced in 1980 include the following:

- There were delays in the release of some of the census data products, particularly those containing data from the sample or long-form questionnaire.

- Maps and other geographic materials were produced in separate clerical operations, leading to delays and inaccuracies that we had to correct before releasing the data products.
- Many of the temporary census offices experienced problems in hiring and retaining workers.
- There were concerns about the accuracy of the 1980 census that led to several legal challenges. In some cases, these challenges forced temporary census offices to remain open longer than planned.

Many major decisions and choices will have to be made in the next year or two. In most cases there will not be one right answer nor a perfect solution. In these cases, we will have to strike a proper balance between competing alternatives. In making these choices, we will consider a number of criteria:

- First, we must meet our constitutional and legal mandates to deliver apportionment counts to the President by December 31, 1990, and the counts for redistricting to the states by April 1, 1991. Any changes in the census process for 1990 must enhance the Census Bureau's ability to meet these deadlines.
- Second, we want to produce all data products from the 1990 census in a more timely manner than ever before.
- Third, we want to keep the total cost of the census reasonable. The aim for 1990 is to keep the per-unit cost (adjusted for inflation) no higher than it was for 1980.
- Fourth, we want to take a census that is as accurate as possible. This is a real challenge: we want to make the census faster and keep costs reasonable, but without reducing the accuracy of the census data.
- Fifth, in deciding what questions we will ask, we must strike a proper balance between the need for information and the time it takes respondents to complete the questionnaire. There will be more legitimate demands for

data, but we must keep the length of the questionnaire reasonable while meeting basic data needs.

- Sixth, we must maintain the strictest confidentiality of each respondent's answers. The success of the census depends directly upon the willingness of the public to cooperate, and their trust in our pledge of confidentiality is one basis for that willingness.

The following will discuss some of the major issues confronting us as we plan the 1990 census -- basic procedures, automation, personnel, outreach and promotion, and questionnaire content.

### Basic Procedures

First, let's look at the basic procedures we will use to take the census. As in 1980, we plan to use the basic mail-out/mail-back method for most of the country in 1990. This involves mailing questionnaires to every household on our mailing list (after taking great effort to assure that the list is as complete as possible), asking householders to fill out their questionnaires and mail them back, and contacting only those housing units for which questionnaires are not returned or for which additional information is needed. In sparsely populated areas of the country where mail procedures are not appropriate we will visit every housing unit. In all areas, we must devise appropriate procedures for enumerating special places (such as college dormitories, military barracks, and prisons), and we will implement quality checks and coverage-improvement procedures to make the census as accurate as possible.

For the 1990 census, we are investigating possible modifications or refinements to the procedures used in the 1980 census. For example, in our 1985 test census in Jersey City, New Jersey, we tested a two-stage census approach. Under this approach, we collected basic information first and two months later collected the additional information from a sample of persons. In 1980, we collected both types of information at one time. Some believed that the two-stage approach might improve the census in hard-to-enumerate urban areas. However, the speculated advantages for the two-stage approach did not materialize. The mail-return rate for the second stage was so low (15 percent) that we discontinued followup for this stage. We do not plan to test a two-stage census further before 1990.

In the 1985 test census in Tampa, Florida, we tested mail-reminder cards. We wanted to see

whether we could improve mail-return rates and reduce costly personal visits by sending reminder cards a few days after questionnaire mailout to households that had not yet returned their questionnaires. This test indicated that reminder cards can be cost-effective, and we will use them again in our 1986 tests.

We will continue to refine census procedures in our 1986 tests. In our test census in part of Los Angeles County, California, we will examine ways to minimize problems caused by mail-delivery in multiunit apartment buildings where apartments sometimes are not well defined. In our test census in several counties in East Central Mississippi, we will work on problems related to getting questionnaires to the correct households in rural-route delivery areas. One option we will test in Mississippi is having census enumerators, rather than the Postal Service, deliver the questionnaires and update the mailing list at the same time. We also will examine some of our enumeration procedures for American Indian reservations in Mississippi.

Special coverage-improvement procedures are an important part of taking a good census. In 1980, these included several operations to improve our address list, a recheck of "vacant" units to see if they were occupied, list-matching, a "were you counted" campaign, and so on. For 1990, we will evaluate the 1980 census coverage-improvement procedures to see which should be repeated, and we will test refinements in these procedures.

One coverage-improvement procedure we are giving special attention to is the Local Review Program. In 1980, for the first time, we gave local officials in over 39,000 jurisdictions an opportunity to review census counts before the temporary census offices closed. Local officials noted any discrepancies between these counts and their own data, and we checked the counts and made corrections, as necessary. For 1990, we want to improve this program. We are working on a design that will give local officials an opportunity to review address counts before Census Day and actual census field counts before the offices close. They would have more time to prepare their data and review our counts. We expect to begin contacting local officials earlier than in 1980, and we are considering holding training sessions, in cooperation with state agencies, to help the localities get ready for the program.

In addition to improving census procedures in order to make the census more accurate, we will continue to examine different undercount measurement and adjustment techniques to

determine whether we can develop a valid procedure for adjusting the census counts. Methods of making adjustments to the counts must not only be statistically sound, they must be legally and politically acceptable. They also must be practical to implement in time to meet our legal deadlines. The 1986 census in Los Angeles County will test the feasibility of quickly measuring coverage so that we could adjust data, if necessary, in a timely manner.

### Automation

Another major issue we are investigating for 1990 is automation. I will just mention automation briefly here since it is addressed in more detail in other papers for this conference.

Basically, our aim is to increase the use of automation in the 1990 census in order to take the census more quickly, more cost-efficiently, and more accurately. Traditionally, census data collection and much of the census data processing have been paper- and people-intensive tasks. The use of automated equipment can help us reduce the mountains of paper and the thousands of clerical tasks and to deal with the whole census process in a much more efficient and controlled way.

Increasing automation in the census will involve two approaches. The first is to automate many of the census tasks performed clerically in 1980 and previous censuses. These tasks include mapmaking, address list updates, questionnaire check-in and editing, coding of written entries on the questionnaires, and cost and progress reporting.

The second approach is to begin automated data processing earlier than in 1980. In the 1980 census, there was a sharp division between the data collection and data processing phases. We did not begin data processing until all the work in a particular district office was completed -- usually 5-7 months after Census Day. For 1990, we will begin processing simultaneously with data collection.

Although we have decided to do the processing earlier, we must still decide where to do it (i.e. in a few centralized or many decentralized locations) and how to do it (i.e. using film-to-tape, keying, or optical mark readers). We must make these major decisions at least by September 1986 to begin the lengthy process of procuring equipment. We will try to resolve as many of the issues as possible even earlier than that and will be addressing them at a "decision" conference in mid-October 1985.

### Personnel

While designing a workable census system is important, we must also have a good work force to get the job done. In some areas in 1980, we had problems hiring and retaining enough good census workers. This was due in part to our pay rates (which may not have been competitive in all areas), the temporary nature of the jobs, and the fact that census work, particularly the personal visits to nonresponding households, can be very difficult. We are giving special attention to finding new ways to recruit, hire, and retain our temporary census work force for 1990.

In our 1985 test census in Jersey City we introduced part-time work on a limited basis. In 1980, we discouraged part-time workers, preferring instead to hire only those who were available to work a 40-hour week. While we still prefer full-time workers, we realize that many qualified people may be able to work only on a part-time basis. In future test censuses, we will examine different methods of paying our enumerators, such as hourly rates, piece rates, performance bonuses, and cost reimbursement. Still, we must provide additional types of nonmonetary motivation. Along these lines, we will investigate job enrichment efforts that would allow temporary workers to see and participate in more tasks. We will also consider new strategies to recruit more motivated and skilled people by seeking active support from community, nonprofit, civic, and volunteer groups.

### Outreach and Promotion

Without public cooperation, it would not be possible to conduct an accurate census in a cost-efficient and timely manner. In past censuses, we have achieved extraordinary levels of cooperation by emphasizing in our promotion campaigns the importance of the census and the strict confidentiality of individual census responses. An effective promotion campaign increases mail-return rates and, thus, reduces the amount of costly followup work. For 1990, we are estimating savings of \$5-6 million for each one (1) percentage point increase in the mail-return rate.

The centerpiece of the 1980 census promotion program was the advertising effort developed by an agency chosen by the Advertising Council. This advertising program provided \$38 million worth of "air" time at no cost to the Census Bureau (except for basic administrative charges). We have asked the Ad Council to undertake another advertising campaign for the 1990 census, and they have agreed to do so. This early decision will allow the Ad Council to be involved

in our test censuses, where promotion ideas can be developed and fine-tuned.

In addition to the advertising campaign, we will have a number of joint ventures with local officials. In the Local Review Program, local officials will be given an opportunity to review precensus housing unit counts and postcensus housing and population counts and to note any discrepancies between these counts and their own data. The Census Bureau will then check any discrepancies in the field, if necessary, and make corrections.

Second, we will again ask communities to set up Complete Count Committees. More than 4,000 of these committees were established in the 1980 census to help generate support for the census using local resources. We feel they were very useful to our outreach efforts, and we want to make this program even better in 1990. In our 1985 test census in Tampa, the Complete Count Committee, formed by the Mayor, was instrumental in getting local celebrities, such as coaches and players from Tampa sports teams, to appear in public service announcements. The Mayor, himself, also appeared in a television announcement for us.

Finally, we will engage in a number of other grass-roots outreach activities. Our regional office staff have regular contacts with a broad array of community groups, and these contacts will increase and intensify as we approach the 1990 census. Another grass-roots activity for 1990 will be our school project in which we will work with school administrators to teach students about the importance of the census and how to answer the census questionnaire. We held a conference on our school project in July 1985 to discuss with educators from around the country how best to design the program for 1990. Finally, we will ask religious leaders and organizations to encourage their constituents to support the census.

### Questionnaire Content

The final issue I will discuss is the census questionnaire. Since the purpose of the census is to meet data needs for at least a decade, a major part of census planning is selecting the census questionnaire content.

As we consult with data users, we are hearing many more requests for data than we can reasonably satisfy. Most of these requests reflect legitimate needs for a wide variety of data to describe our complex society. For example, we have recently received dozens of letters from all

around the country advocating questions on pets. The advocates make a good case for the importance of pet data in public health planning and animal control programs.

But, one of our goals for the 1990 census is to balance the needs for information against the length of the questionnaires. This balance is necessary because public cooperation is essential for a successful census. Such cooperation could be undermined by questionnaires that the public finds too lengthy. In practical terms, this means that there can be no significant growth in the size of the questionnaires for the 1990 census.

In making the final choices about which subjects to include in the questionnaires, we will follow seven standards:

- First, we will collect only required data -- those needed for constitutional or legislative reasons, those needed specifically to administer Federal, state, and local programs, and those needed to describe the most important aspects of the American population and housing stock.
- Second, the census must meet small-area data needs. If the data are needed for small geographic areas (for example, census tracts with an average population size of 4,000), then the census is an appropriate tool. If the data are required only for larger areas (such as the Nation, regions, states, and large metropolitan statistical areas), sample surveys might be more appropriate.
- Third, we will consider the need to collect data for small and dispersed population groups. The census is more appropriate for this purpose than a nationwide sample survey because a survey would not give adequate coverage of these groups and, thus, would not provide statistically significant data about them.
- Fourth, the questions must lend themselves to self-response. The questions generally will be answered directly by respondents without an enumerator present. So, they must be easy to understand with terminology widely accepted by the public.
- Fifth, the questions must not impose unrealistic requirements for data processing; and the responses must be translatable, with reasonable efforts, to machine-readable form. One thing this means is that the number of write-in questions should be minimized.

- Sixth, we will not consider any question that we believe is intrusive, offensive, or widely controversial. The Census Bureau needs public cooperation for the census to work. It cannot risk losing that cooperation through improper questions.
- Seventh, many of the subject areas to be asked in 1990 will have been asked in 1980 and earlier censuses. Answers in 1990 to questions asked previously can provide trend data needed to analyze vital socioeconomic and housing characteristics. This criterion does not mean that just because we asked a question in the last census, it will be asked again or that we will not ask new questions. We will consider, however, the need to provide continuity and comparability between data gathered during each census.

This month, we have completed our series of local public meetings where data users from across the country advised us on questionnaire data items, data products, and census geographic areas. We held at least one meeting in each state. We also have completed meetings with the Federal agencies to determine which data they need to administer Federal programs, and we are analyzing the results of that process.

We must make the many decisions about census content in the next 2-3 years. Indeed, we are now planning our National Content Test for 1986, our main vehicle for testing new questions and question wordings. By law, we are obligated to report to the Congress on the subject areas for the census by April 1, 1987, and on the actual questions that we will ask by April 1, 1988.

## Closing

Even without having to plan for change and to make needed improvements, taking a census is an exciting and tremendous challenge. Enumerating and collecting detailed characteristics for over 226 million people and 88 million housing units, as we did in 1980, were not simple tasks. This is due in part to the highly mobile nature of people in the United States, and the diverse conditions and situations in which we live. Our task will be more difficult in 1990, when we estimate that there will be about 24 million more people and about 18 million more housing units.

Adding to the challenge is the fact that we must count not only the majority of us who live in houses, apartments, condominiums, trailers, and so on, but also those who live in group quarters -- such as military barracks, college dormitories, penal institutions, long-term hospitals, and migrant farm camps -- and even those without any home.

This paper has outlined some of the major issues the U.S. Census Bureau faces in planning the 1990 Census of Population and Housing. While these five -- basic procedures, automation, personnel, outreach and promotion, and questionnaire content -- are some of the major issues, there are thousands of other issues and decisions that make taking the census complex and challenging.

We look forward to discussing these issues and sharing problems and solutions with our colleagues at this conference. We also look forward to learning about the key issues being faced by the other participating countries as they plan their next censuses.



# PLANS FOR FUTURE AUSTRALIAN POPULATION CENSUSES

HENRY KRIEDEL

## POPULATION CENSUS SYSTEMS DEVELOPMENT AND LONG-RANGE PLANNING AUSTRALIAN BUREAU OF THE CENSUS

The Australian Bureau of Statistics (ABS) is planning to address over the next two years a number of issues in respect of the conduct of population censuses after 1986. These issues range from conceptual to methodological, including matters relating to subject content which will obviously need to be addressed in detail at a later stage. The objective of this paper is to discuss the major issues which ABS plans to address.

### Topic Selection

For the 1981 and 1986 Censuses, an important aspect of the selection and development of topics for inclusion in these censuses was a topic submission scheme. This involved inviting submissions on topics from major users and the public.

The submissions were analysed by the ABS, some high priority topics were tested and the final ABS recommendations on topics were based on reactions to announce preliminary ABS views on topics for the census. For 1986, submissions received on certain topics during the 1981 submission scheme were used in the selection of topics.

The 1986 submission scheme revealed no unknown major topics and very little extra insight into user demands and their reasons for requesting topics. A similar scheme for the next census is unlikely to improve on our existing knowledge of user needs. It is considered that the approach needs to change, while still maintaining the important element of public consultation and some means of establishing new needs. What would be more productive is a greater analysis of user need for the various known potential census topics with more attention paid to trends in data collected in the past to determine better the need for the data to be collected every five years or by survey. Consideration should also be given to linking the assessment to the planning for household surveys.

Development of topics could be based on the simultaneous publication of a series of discussion papers on various potential topics, similar to Preliminary Views but with more analysis of past data trends to support views and more attention paid to satisfying needs by either census or surveys. Reactions from users and the public to

these discussion papers, particularly advice on new needs, would form the basis of selection of final recommendations.

### Basis of Enumeration

All Australian censuses to date have been conducted on the basis of enumerating people at their actual location on census night. Until the 1971 Census, all census statistics for an area referred to persons counted in that area on census night. A question on a person's address of usual residence was first included in the 1976 Census primarily for the purpose of measuring internal migration and was used to produce a very limited number of tables on a place of usual residence basis. The question was used in the 1981 Census for producing resident population estimates as well as limited census outputs on a place of usual residence basis.

Addresses of usual residence at census time were coded to census local government area (LGA) in the 1976 and 1981 Censuses and will be coded to statistical local areas (SLAs) for the 1986 Census. This allows the production of tables on a place of usual residence basis for areas which can be formed by aggregation of LGAs (now SLAs). With the adoption of the estimated resident population series since 1981 there has been an increased demand for census data on a place of usual residence basis. For the 1986 Census this will be partly met by the production of a greater range of data on a place of usual residence basis.

There are, however, limits to the extent that census data on a place of usual residence basis can be produced from the current methods. No such data can be produced for areas which cannot be defined by SLAs nor for full characteristics of families and households from which usual members were temporarily absent on census night. To overcome some of the problems with family statistics, a question on usual residents temporarily absent on census night has been included on the 1986 Census households form. Only basic demographic data will be obtained for such persons for use in improving the counts of families by types. Deficiencies in family statistics such as family income cannot be improved by the inclusion of this question.

More areas could have data produced on a place of usual residence basis by coding addresses of usual residence to collection districts (CDs), but at a substantial cost. Overcoming the problems of family and household data would, however, require data relating to persons and families absent from their usual residence but enumerated elsewhere to be transferred to their usual residence. This could be done clerically at the Data Transcription Centre (DTC) but would probably delay processing and the first release of data, and would be very expensive. Alternatively, it could be done by computer if addresses were available (temporary) on the computer record at the cost of capturing the address data. To do so, however, would require a change in the existing policy of not entering addresses onto a computer record. Such a change in policy is unlikely to occur.

Conducting the census on the basis of enumerating people at their usual residence would overcome a number of problems of the current method employed for Australian censuses. This is the basis on which censuses are conducted in the U.S.A., Canada, Japan and most European countries. Changing the basis of enumeration would result in household forms containing full details for the usual residents of that household thus avoiding the current need for extra costs to be incurred during processing. It would, however, cause a break in census time series data used for longitudinal studies and may mean that census counts of persons in an area on census night would no longer be available (an important user need especially for areas with highly transient populations, e.g., holiday resorts). Another aspect of changing the basis of enumeration would be the likelihood of some degree of respondents confusion.

A change of basis would require considerable development costs, even taking into account the experience of other countries. Particular conceptual issues that would need to be resolved include treatment of Australians temporarily overseas on census night and persons with no usual residence.

### Census Statistical Geography

A number of aspects of statistical geography used for production of output from the census require investigation to establish if improvements or modifications are needed.

- (a) The increasing demand for small area data from the census has resulted in more and more use of CDs as the smallest output unit

for data and for aggregating data to user-specified small areas. CDs are primarily designed for the administration of the field work for the census and are not ideal dissemination units. On the one hand they are generally too large to allow accurate aggregation to non-standard larger areas but on the other they are too small to allow for basic detailed data to be produced because of confidentiality constraints.

To better satisfy user needs for more accurate delimitation of their many and varied small area units requires at least the coding of household addresses to block level or ideally to point reference them by the allocation of geocodes. Coding addresses to blocks will require the development and production of census block maps for the whole of Australia consistent with CDs. It may be possible to do this relatively cheaply but the amount of data released to users at block level would be considerably less than currently released for CDs because of confidentiality constraints. The ABS would be able to access more detailed data on a data base with data stored at block level to produce aggregated data which more closely corresponded to user-defined areas. Such a strategy would, however, result in a greatly increased demand for ABS resources to satisfy ad hoc requests, reversing the trend established over recent censuses of encouraging more users to obtain summary files to aggregate data. Greater attention would also need to be paid to ensure that identifiable data are not obtained by subtracting data for closely defined areas to arrive at detailed data for very small areas.

Geocoding of addresses could be done by having collectors mark the location of households onto field maps and the marks digitized by reference to the already digitized CD boundaries. This would, however, effectively result in addresses being entered onto the computer record and therefore would require a change to existing policy.

An alternative approach would be the creation of a new small area output unit larger than CDs to allow for the release of more detailed data but probably no greater in size than 10,000 persons. Such a new unit would be similar in size and purpose to census tracts used in the U.S. Census and should aim to fit into the ASGC (Australian Standard Geographical Classification)<sup>1</sup> hierarchy between CDs and statistical

<sup>1</sup> Note from the editor.

subdivisions. It would, therefore, likely be formed by combining small SLAs and splitting large ones.

- (b) It is 20 years since the existing delimitation criteria for standard area units were established. A number of minor changes have been made to the criteria since they were first decided, particularly for the creation of statistical districts. It is considered that the existing criteria need examining in the light of the existing population distribution and user needs.
- (c) The creation of standard destination zones for place of work data which are consistent with the ASGC would allow the storage of destination zone codes for units on the ABS integrated business register. This could reduce the cost of preparation of coding material for each census and allow the production of intercensal estimates of employment by destination zone (after appropriate allowance for undercoverage of the register).
- (d) The recording of postcode number and suburb reported on the census form as a person's address would allow for data to be more accurately compiled for such areas instead of the approximations made at present by the allocation of whole CDs to such areas. Postcode areas and suburbs are not included in the ASGC (except where individual SLAs equate with individual suburbs) but the demand for census data for such areas continues to increase as more data for these areas are released from administrative systems. Alternatively, this requirement may be able to be satisfied through the point referencing of geocodes, should point referencing be implemented.

As well as these census specific issues, ABS Classification Section intends to proceed with specification, linking and rationalisation (to the extent possible) of all other geographical areas and classifications which the ABS uses in the provision of statistics (such as parishes in the State of Victoria and postcode areas) but which for technical and other reasons could not be incorporated into the ASGC. The results of this work will be incorporated in a comprehensive ABS reference manual on statistical geography entitled the "Geographical Classification Framework".

## Mapping

The Division of National Mapping, which provides the necessary field and dissemination maps for each census, plans to develop an extensive geographical data base, including digitized map data and a system for computer generation of maps. This should reduce the cost of map preparation for each census and allow for quick and flexible generation of all maps required for field and dissemination purposes. The ABS has been invited to participate in these developments.

The creation of an Australia-wide land information data base utilising existing State- and local area-based land information systems will help to keep development costs to a reasonable level. The base would enable the unique identification and listing of dwellings which in the long run has potential use in conducting mail-back censuses, coding persons and households to address of usual residence and for improving the production of census data for user-defined areas. These issues are discussed in other sections of this paper.

## Mail-back Census

Prior to the development of the 1986 Census, an investigation was made into mail-back censuses, primarily as a means of providing greater protection of privacy of census forms. This investigation showed that a mail-back census could be cheaper than the current drop-off and pick-up method if more than 80% of households mailed their forms within the first two weeks after census night. Collector follow-up would be necessary of households not responding or returning significantly incomplete forms. The lower the level of non-response or serious partial response, the greater the savings compared with the current method.

This investigation did not examine whether CDs could be increased in size because more forms could be delivered without the constraint of having to collect the forms which usually takes longer. Larger CDs would require less field staff and hence reduce costs. In practice, it is likely that CDs would be made larger only in growth areas by raising the size criteria to be reached before a CD is split. This is because of the need to maintain comparability with previous censuses and the high cost of establishing a completely new set of CDs.

## Telephone Follow-up

Telephone follow-up of households with significant non-response to questions is undertaken in

the U.S. and Canadian censuses to improve data quality while keeping costs low. The follow-up is mostly done from regional offices controlling field operations but some is also done from the central processing centre.

Greater use of the telephone for follow-up work is being considered by the USBC for their 1990 Census. Such use would be linked to automated mark-in of mailed-back forms which would generate work-loads for operators using computer-assisted telephone interviewing (CATI). Follow-up work in the Australian census is undertaken by collectors at the time of collecting forms. If a mail-back census is introduced, serious consideration must be given to asking for telephone numbers on census forms so that telephone follow-up could be performed to reduce costly field follow-up. Such a method would require testing to determine the public's willingness to supply telephone numbers and to convey information for census questions over the telephone, and the extent of non-contact.

### Processing Developments

Investigations have commenced into alternative data capture methodologies for future censuses. Quite apart from concerns about issues such as repetitive stress injuries (RSI) and job design, key entry as is being used for 1986 is a very expensive and labour-intensive task. A methodology which appears to offer an attractive solution is a respondent-marked questionnaire containing as many optical mark recognition (OMR) responses as possible. Remaining questions with text answers could be office coded and marked before the forms are read by OMR.

Alternatively, if it is possible to design a questionnaire with, as the minimum, all questions relating to preliminary data for population estimates being covered by OMR responses, it may be possible to avoid the high cost of preliminary processing of the data for purposes

of producing timely revisions to population estimates. All self-coding questions on the forms would be read by OMR to create a file of partially captured records (that is, deficient in that text responses have not yet been captured) which could be processed to produce preliminary census counts required for population estimates. Subsequently, the partially captured records could be individually processed by clerks at key-entry stations. The clerks would view the text responses from the questionnaires, microfilm, or possibly even a video image captured during the OMR process, and would undertake interactive data entry of text fields, coding and/or editing. Variants on this theme have been used in recent French censuses, and are being developed for the 1986 New Zealand Census and the 1990 U.S.A. Census.

Use of automatic coding and/or interactive computer-assisted coding will be investigated. As well as assisting in ensuring uniformity in the coding operation it would assist in alleviating the labour-intensive task of clerical coding. This could form part of the overall processing methodology outlined in the previous paragraph, as is the case in the French, New Zealand and U.S.A. methodologies.

Another issue relating to census processing is the degree of centralised versus decentralised operations. The data transcription centre established for recent censuses, including 1986, has been centralised with significant economies of scale and providing uniform data quality. However, the parameters in the cost equation are changing quite significantly (e.g., hardware, communications networks, recruitment demand at a single location) and it may prove advantageous to consider some degree of decentralisation of computer and/or clerical operations. The problem of ensuring uniform quality of coding which has been experienced by many countries with multiple processing centres could be largely overcome by providing automatic or computer-assisted coding facilities.

## Résumé of Discussion

The second part of the session on the role of research and testing in the census cycle was reserved for a discussion entitled "Where do we go from here?" However, before addressing this specific question, a few questions were taken from the audience.

### 1. Questions from the Audience

One of the questions raised was the various problems different countries had towards the collection of names and addresses and the retention of census documents. All countries recognized the need to collect names and addresses for operational reasons. However, policies seem to vary widely in terms of how long and in what form the names and addresses are kept. Australia, for example, shreds all documents within two years after census. Canada and the U.K. retain the documents for many years, largely for purposes of future historical research, but have a policy of not putting the name or address on the computer. It appears that the possibilities offered by technology, for example, the formation of address registers or the use of record linkage, are often ahead of the public's willingness to accept it.

A second question raised concerned the independence of the vehicles used to measure coverage and to evaluate content. Because content variables (e.g., age and sex) are often required for matching purposes and because they are subject to errors, using the same vehicle to measure both could confound the measurement of the two. The point was also raised that using a post-enumeration survey to measure coverage error may result in an underestimation if the same people tend to be missed. A third question concerned the use of household surveys to test censuses. In Australia, monthly surveys are used in the topic selection process. However, for most of the census-taking countries, the use of surveys for census testing is not frequent due to the differences in collection procedures.

Finally, the panelists briefly exchanged ideas on the costs of testing compared to the total census cost. For most countries, testing seems to represent about 2 to 10% of the total cost of a census.

### 2. Panel Discussion: Where do We Go from Here?

All panelists agreed that research and testing were important in the planning of a census.

However, the problems of mounting programs in the various countries were different because of the different census cycles. **Mr. Bounpane** began the discussion by describing the approach used to plan the U.S. program of testing.

The testing program for the U.S. was developed using a three-fold approach. First, the 1980 Census was examined for problems that should be solved. Second, a series of internal committees was set up to look at a specific topic, in terms of how things were done in the past and how they would be done in the future. Each committee produced a report. Third, an effort was made to obtain outside suggestions about the census.

This process created a huge shopping list of topics which then had to be reduced. This was done, and could only be done, by the professional judgement of the Census Bureau staff. Topics were chosen for testing which involved the most changes for the census and the most potential risk. Other topics were not chosen if it was felt that they could be made to work without testing.

The U.S. also attempted to build support within the Bureau by using a working group of stakeholders. This group is composed of members of different fields. It serves to filter the alternatives, the objectives and make consolidated recommendations on topics to be tested. This group helps to develop a certain appreciation within the Bureau about the involvement of all areas in the census.

Following this, **Mr. E. T. Pryor** gave a brief overview of areas that should be considered for testing for the 1991 Census of Canada.

First, the content is a fundamental area to consult and to test. This is an area that has an overall effect on the census. Content has to be determined early in the census process and should be one of the first areas to be tested.

Second, automation is another area that needs testing. Areas such as planning, collection, MIS systems, and data processing are all areas that have potential benefits from automation.

Third, there is a need to create an environment favourable to the personal development of census staff. How do we recruit, train and motivate our staff in the face of budget constraints and a constantly limited work-force?

Fourth, there is a need to do research and plan a strategy to deal with the issues of privacy and confidentiality.

*Next are collection methods, which encompass things like address registers, extension of mail-back, bilingual questionnaires, centralized edit and sampling.*

*Quality measurement, such as under and overcoverage and content evaluation also need to be researched and tested.*

*Finally, we need to examine the census output, in terms of things like the product line, new technologies and decentralized dissemination.*

*In terms of how we achieve these things, we obviously must have a structure in place and not lose the momentum from this conference.*

### **3. Concluding Remarks by Panelists**

*Panelists took the last minutes of the conference to exchange final remarks and recommendations.*

*Three general remarks were made about censuses. The first one concerned the future. It is important in planning the 1991 Census to*

*keep in mind future censuses. Second, we should remember that history tends to swing back and forth, and what is considered now as a problem may not be a problem in the future. Third, all countries emphasized the importance of the census to the country, and noted that a census could not afford to fail. Changes to a census must therefore be made cautiously.*

*One recommendation was unanimously shared by every panelist. Census-taking countries should develop more exchanges in many ways. Exchanges of staff should be encouraged, especially people who would participate in a testing part of a program or to one particular aspect during the census process. Also, exchanges of information, documentations and experiences should be extended and should become more automatic.*

*Experiences of other countries on testing should be more frequent, as they would be beneficial to the development of census in each country. Because it is difficult to obtain funds and time to do testing, countries should take every advantage of experiences from others.*

## CLOSING REMARKS

Edward T. Pryor

Director General for Census  
and Demographic Statistics  
Statistics Canada

I would like to take advantage of the closing remarks to sincerely thank everybody involved in this conference.

First, I would like to thank our visitors from other countries for their participation, their presentations and their interventions. Their perspectives brought us a lot of new ideas and new approaches on general and specific issues.

Secondly, I would like to thank Mr. Ivan Fellegi and Mr. Bruce Petrie for the support they gave me for the organization of the conference. Without their support the conference would probably not have become a real project.

Of course, the person that I want to thank the most is Mr. Don Royce who was in charge of the arrangements for the conference. Without his dedication to the preparations, it would have been almost impossible to do it.

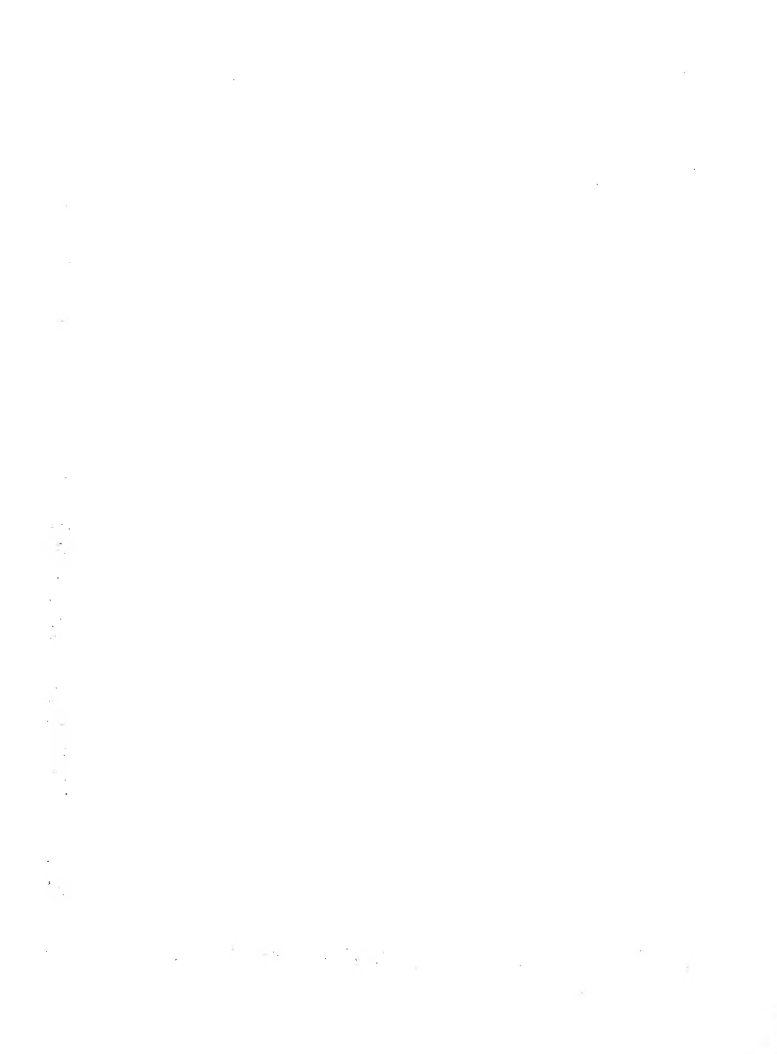
I also would like to thank the program committee composed of Ken Brown, Richard Ethier, Gilles Montigny, Henry Puderer, Doug Hicks and Wilson Freeman. This committee developed the program that was so easy to follow during the conference.

I would also like to thank the local arrangement committee of which members were Kevin Thatcher, Karen Kelly, Kathy Miller, Owen Power, Anis Ashraf and Josée Dufresne. This committee was responsible for a large part of the success of the conference.

My final thanks are reserved for the participants for their enthusiastic participation to the conference. Because of them, Statistics Canada now has a basis to launch the 1991 program. This program will be built from ideas and approaches that were suggested so generously by you. Thank you.



## APPENDICES





- 14:00-17:00      Session: **Census Geography**
- Chairperson:            D. Ross Bradley  
Director,  
Geography Division
- 14:00            Geographic Support for the 1990 Decennial Census  
Silla Tomasi, Assistant Division Chief for Operations, Geography Division, U.S.  
Bureau of the Census
- 14:50            Coffee
- 15:10            Area Master Files - A Better Way to Serve Census Needs  
Joel Yan, Geocartographics Sub-Division
- A Block Program - Yes or No  
Robert Parenteau, Geography Division
- Computer Systems to Support Census Geography  
Gordon Deecker, Geocartographics Sub-Division
- 16:15            Participant:            S. Witiuk, Assistant Director, Informatics Services and  
Development Division
- 16:25            Open Discussion

**Wednesday, October 9**

- 9:00-12:00      Session: **The Content of the 1991 Census**
- Chairperson:            Ian Macredie  
Director,  
Labour and Household Surveys Analysis Division
- 9:00            Discussion:            The 1991 Content Development Process
- Panelists:            Henry Kriegel, Director, Population Census Systems  
Development and Long-range Planning, Australian  
Bureau of Statistics
- Susan Miskura, Chief, Decennial Planning Division,  
U.S. Bureau of the Census
- David Pearce, Head, Census Division, Office of  
Population Censuses and Surveys, United Kingdom
- 10:00            Coffee
- 10:15            Discussion:            1991 Content Proposals
- Speakers:            John Kettle, Futuresearch Publishing Inc.
- Frank Clayton, Clayton Research Associates
- Noah Meltz,  
University of Toronto
- 11:45            Open Discussion

- 13:30-16:30      Session: **The Census of Agriculture**
- Chairperson:            Terry Gigantes  
Director General,  
Resources, Technology and Services Statistics Branch
- 13:30            Mail Enumeration in the U.S. Census of Agriculture  
Charles Pautler, Chief, Agriculture Division, U.S. Bureau of the Census
- Proposal for a Land-based Census of Agriculture  
Oliver Code, Agriculture and Natural Resources Division
- 14:30            Coffee
- 15:00            Confidentiality Procedures for the 1991 Census of Agriculture  
Rick Burroughs, Agriculture and Natural Resources Division  
Mary March, Census and Household Survey Methods Division
- 15:30            Participant:            Mel Jones, Census Manager, 1986 Census of  
Agriculture
- 16:00            Open Discussion

**Thursday, October 10**

- 9:00-12:00      Session: **Census Automation**
- Chairperson:            Martin Podehl  
Director,  
Informatics Services and Development Division
- 9:00            Automation Plans for the 1990 U.S. Census of Population and Housing  
Peter A. Bounpane, Assistant Director for Demographic Censuses, U.S. Bureau  
of the Census
- Data Capture Methods - The Alternatives  
Dave Croot, Client Services Division
- Decentralized Data Capture Methods for the U.S. Census  
Arnold Jackson, Chief, Decennial Operations Division, U.S. Bureau of the  
Census
- 10:30            Coffee
- 10:45            Automated Coding at Statistics Sweden  
Lars Lyberg, Statistics Sweden
- Generalized Software  
T. Mike Jeays, Informatics Services and Development Division
- 11:30            Participant:            J. Ryten, Assistant Chief Statistician, Informatics and  
Methodology Field, Statistics Canada
- 11:45            Open Discussion

- 13:30-16:30      Session: Coverage and Data Quality
- Chairperson:            Gordon Brackstone  
Director General,  
Methodology Branch
- 13:30            Issues in Coverage Measurement and Adjustment  
Howard Hogan, Chief, Undercount Research Staff, Statistical Research Division,  
U.S. Bureau of the Census
- Making Data Quality Assessment More Relevant  
Richard Burgess, Census and Household Survey Methods Division
- Adjustment for Non-coverage Errors  
Chris Hill, Census and Household Survey Methods Division
- 15:00            Coffee
- 15:20            Address Registers, Advantages and Disadvantages  
David Whitford, Chief, Research Co-ordination Branch, Decennial Planning  
Division, U.S. Bureau of the Census
- Applications of Address Registers in the Canadian Census  
Don Royce, Census and Household Survey Methods Division
- 16:00            Open Discussion

# **Friday, October 11**

- 9:00-11:50      Session: The Role of Research and Testing
- Chairperson:            E.T. Pryor  
Director General,  
Census and Demographic Statistics Branch
- 9:00            Discussion:            Planning the 1990 Round of Censuses – An  
International Perspective
- Speakers:            David Pearce, Head, Census Division, Office of  
Population Censuses and Surveys, United Kingdom
- Peter Bounpane, Assistant Director for Demographic  
Censuses, U.S. Bureau of the Census
- Henry Kriegel, Director, Population Census Systems  
Development and Long-range Planning, Australian  
Bureau of the Census
- 10:30            Coffee
- 10:50            Discussion:            Where do we go from here?
- Panelists:            David Pearce, Peter Bounpane, Henry Kriegel and  
Edward Pryor
- 11:20            Open Discussion
- 11:50            **CLOSING REMARKS**

## APPENDIX 2 - LIST OF PARTICIPANTS

Barnabé, Richard  
Regional Director  
Quebec  
Regional Operations Branch

Bradley, Ross  
Director  
Geography Division  
JT-3B7

Clayton, Frank  
Clayton Research  
Associates

Cunningham, Ron  
Informatics Services and  
Development Division

Gigantes, Terry  
Director General  
Resources, Technology and  
Services Statistics Branch  
JT-13B6

Hogan, Howard  
Chief  
Undercount Research  
Staff, Statistical  
Research Division  
U.S. Bureau of the Census

Jones, Mel  
Census Manager  
Census of Agriculture  
Agriculture and Natural  
Resources Division  
SC-3000

Bounpane, Peter A.  
Assistant Director for  
Demographic Censuses  
U.S. Bureau of the Census

Burgess, Richard  
Chief  
Data Quality and Analysis  
Section  
Census and Household Survey  
Methods Division  
JT-4C6

Code, Oliver G.  
Chief  
Crops Section  
Agriculture and Natural  
Resources Division  
SC-2401

Deecker, Gordon  
Chief  
Geocartographics Centre  
Informatics Services and  
Development Division  
JT-2A2

Hicks, Doug  
Chief  
Census Collection Operations  
Survey Operations Division  
JT-6C7

Jackson, Arnold A.  
Chief  
Decennial Operations Division  
U.S. Bureau of the Census

Kazmaier, John A. Jr.  
Assistant Division Chief for  
Censuses  
Field Division  
U.S. Bureau of the Census

Brackstone, Gordon  
Director General  
Methodology Branch  
JT-5B8

Burroughs, Rick  
User Services  
Census of Agriculture  
Agriculture and Natural  
Resources Division  
SC-3000

Croot, Dave A.  
Director  
Client Services Division  
SC-2401

Fellegi, Ivan P.  
Chief Statistician  
R.H. Coats 26-A

Hill, Chris  
Chief  
National Task Force on  
Tourism Data  
R.H. Coats 11-C

Jeays, Mike  
Assistant Director  
Research and General Systems  
Sub-Division  
Informatics Services and  
Development Division  
R.H. Coats 13-A

Kettle, John  
Futuresearch  
Publishing, Inc.

Kidd, Karole

Informatics Services and  
Development Division

Kriegel, Henry

Director  
Population Census  
Systems Development and  
Long-range Planning  
Australian Bureau of  
Statistics

Lyberg, Lars

Statistical Research Unit  
Statistics Sweden

Macredie, Ian

Director  
Labour and Household  
Surveys Analysis Division  
JT-6A8

March, Mary

Senior Methodologist  
Census and Household Survey  
Methods Division  
JT-4B6

Meltz, Noah

University of Toronto

Miskura, Susan

Chief  
Decennial Planning  
U.S. Bureau of the Census

Page, Jerry C.

Regional Director  
Alberta, Northern  
Saskatchewan and N.W.T.  
Regional Operations Branch

Parenteau, Robert

Officer  
Spatial Delineation and  
Analysis Section  
Geography Division  
JT-3B7

Parker, Jean-Pierre

Geocoding Base File  
Spatial Systems Section  
Geography Division  
JT-3A6

Pautler, Charles P.

Chief  
Agriculture Division  
U.S. Bureau of the Census

Pearce, David

Head  
Census Division  
Office of Population Censuses  
and Surveys  
United Kingdom

Podehl, Martin

Director  
Informatics Services and  
Development Division  
R.H. Coats 13-A

Pryor, Edward T.

Director General  
Census and Demographic  
Statistics Branch  
JT-5B8

Riddle, John

Associate Director General  
Regional Operations Branch  
JT-6C8

Royce, Don

Chief  
Operations Section  
Census and Household Survey  
Methods Division  
JT-4B5

Ryten, Jacob

Assistant Chief Statistician  
Informatics and Methodology  
Field  
JT-13B8

Tomasi, Silla G.

Assistant Division  
Chief for Operations  
Geography Division  
U.S. Bureau of the Census

Underhay, Boyd J.

Regional Director  
Newfoundland and Labrador  
Regional Operations Branch

Whitford, David C.

Chief  
Research Coordination Branch  
Decennial Planning Division  
U.S. Bureau of the Census

Williams, Brian J.

Assistant Regional Director  
Manitoba and Southern  
Saskatchewan Regional Office  
Regional Operations Branch

Witiuk, Sid

Assistant Director  
Geocartographics Sub-  
Division  
Informatics Services and  
Development Division  
JT-2A5

Yan, Joel

Chief  
Methodology Geocartographics  
Sub-Division  
Informatics Services and  
Development Division  
JT-2A2



STATISTICS CANADA LIBRARY  
BIBLIOTHEQUE STATISTIQUE CANADA



1010232695

c.4



